ErrorIST 2.0 - Geração de erros *à la carte*

Raquel Marçal Cristóvão

Dissertação para obtenção do Grau de Mestre em

Engenharia Informática e de Computadores

Orientadores: Prof. Maria Luísa Torres Ribeiro Marques da Silva Coheur
Drª Helena Gorete Silva Moniz

Júri

Presidente: Prof. António Manuel Ferreira Rito da Silva
Orientador: Prof. Maria Luísa Torres Ribeiro Marques da Silva Coheur
Vogal: Prof. João Paulo Baptista de Carvalho

Outubro 2018

Agradecimentos

Durante todo o processo de desenvolvimento da tese, apesar de ser um trabalho individual, tive contributos de várias pessoas. Primeiro que tudo e todos, gostaria de agradecer aos meus pais e irmã, que me deram a oportunidade de tirar um mestrado e que me apoiaram e encaminharam, sempre, da melhor forma.

Ao longo de todo o percurso escolar, tive a oportunidade de desenvolver projetos com pessoas que me transmitiram muitos conhecimentos e que se tornaram importantes sem me dar conta.

Obrigada à Professora Luísa Coheur e à Professora Helena Moniz, orientadoras da dissertação, que se disponibilizaram desde o início para me ajudar e contribuíram incansavelmente para este trabalho. Quero, também, agradecer a todas as pessoas que dedicaram parte do seu tempo à realização da avaliação do trabalho.

Por último, agradeço ao Instituto Superior Técnico - Taguspark, pelos desafios diários, pelo crescimento intelectual e emocional e, ainda, pelas condições de estudo proporcionadas ao longo destes anos.

Abstract

The evaluation of learners of a language is a task that requires time and work, and that brings an increase in cost to evaluators. But what if evaluators could automatically generate errors and evaluate learners?

In this document, we present ErrorIST 2.0, errors *à la carte*, a tool capable of generating and inserting errors in texts, and evaluate the human interventions. The implemented tool offers several types of errors – such as syntactic, morphologic, punctuation and spelling errors – that can be adapted to different evaluation contexts, such as evaluation of students or editors.

The system generates errors, by taking into account the needs of the evaluator, and inserts them into texts. After the intervention of the students and/or editors, the system detects the corrections and assigns them a classification.

ErrorIST 2.0 allows the insertion of generic errors and, also, of more language specific errors. Although having Portuguese as the main language, in this thesis, we also tested ErrorIST 2.0 in English and French.

The process of evaluating the corrections automatically is challenging and although ErrorIST 2.0 cannot replace the human intervention, it reduces it in more than 45%, depending on the evaluation scenario.

Keywords: errors, taxonomy, error generator, ErrorIST, evaluation errors

Resumo

A avaliação de aprendentes de uma língua é uma tarefa que exige tempo e trabalho e que traz custos aos avaliadores. Mas e se os avaliadores pudessem gerar erros automaticamente e avaliar aprendentes?

Neste documento, apresentamos o ErrorIST 2.0 *à la carte*, uma ferramenta capaz de gerar e inserir erros em textos e, ainda, avaliar as intervenções humanas. A ferramenta implementada oferece diferentes tipos de erros – sintáticos, morfológicos, pontuação e ortográticos – que podem ser adaptados a diferentes contextos de avaliação, como a avaliação com estudantes ou editores.

O sistema gera erros, tendo em conta as necessidades do avaliador, e insere-os em textos. Depois das intervenções dos estudantes e/ou editores, o sistema deteta as correções e atribui-lhes uma classificação.

O ErrorIST 2.0 permite a inserção de erros genéricos e, também, de erros mais específicos de uma dada língua. Embora tenha como idioma principal o português, nesta tese, iremos testar o ErrorIST 2.0 em inglês e francês.

O processo de avaliação das correções de forma automática é um desafio e embora o ErrorIST 2.0 não consiga substituir a intervenção humana, consegue reduzi-la em mais de 45%, dependendo do cenário de avaliação.

Palavras-chave: erros, taxonomia, geração de erros, ErrorIST, avaliação de erros

Conteúdo

Lista de Figuras

Lista de Tabelas

Capítulo 1

Introdução

1.1 Motivação

A avaliação de fluentes ou aprendentes de uma língua é uma das preocupações para alguns projetos na área de Processamento de Língua Natural. Com a utilização do ErrorIST 1.0, uma ferramenta desenvolvida no âmbito da tese de mestrado (dos Santos, 2016), é possível gerar, inserir diferentes tipos de erros em textos e, por comparação de textos, avaliar automaticamente o falante. Este documento descreve uma nova versão do ErrorIST 1.0, o ErrorIST 2.0, uma ferramenta dedicada à geração de erros. Dado um texto, o ErrorIST gera um conjunto de erros – sintaxe, morfologia, pontuação, ortografia e semântica – e adiciona-os ao texto fornecido de acordo com a escolha dos tipos de erros feita pelo avaliador. Uma vez introduzidos os erros, o texto é submetido ao avaliado para correção. Após a correção, o ErrorIST avalia o desempenho do avaliado de acordo com as suas intervenções. Para além da área empresarial, temos outros cenários possíveis para a utilização do ErrorIST 2.0 como a avaliação da destreza de alunos e estrangeiros numa determinada língua.

A nossa motivação neste trabalho é desenvolver uma versão melhorada da ferramenta ErrorIST 1.0, disponibilizando novos tipos de erros, novos idiomas e conseguir adaptar o sistema aos vários cenários de avaliação, evitando a geração de erros manual e grande parte da correção.

1.2 Objetivos

A criação do ErrorIST 2.0 tem como propósito melhorar e desenvolver o trabalho anteriormente apresentado. Após uma análise do ErrorIST 1.0, identificámos os principais pontos a melhorar e delineámos os seguintes objetivos para o ErrorIST 2.0:

- Melhorar o processo de geração dos erros relacionados com o grafar de maiúscula e mninúscula, pontuação e conjugação de verbos;
- Aumentar o número de erros disponíveis, inserindo novos tipos de erros de sintaxe, morfologia e ortografia;

– Integrar erros de inconsistência no discurso ao longo de um documento, como os erros de variação dialetal na língua e os erros de registo;

– Mostrar a usabilidade da ferramenta em novos idiomas, como o inglês e o francês;

– Rever o processo de avaliação e avaliar o ErrorIST 2.0 tendo em conta três potenciais utilizadores: alunos de uma disciplina de escrita; fluentes em francês; pós-editores da empresa de tradução Unbabel.

1.3 Contribuição

O ErrorIST 2.0 começa por oferecer aos utilizadores um cardápio de erros que não se foca apenas numa taxonomia e que têm a capacidade de se adaptar ao tipo de avaliação desde os erros mais genéricos aos mais específicos de cada idioma. Uma das funcionalidades adicionais desta versão é a possibilidade de verificar a existência de uma palavra gerada através de um novo módulo de verificação, evitando que sejam geradas palavras que não constem no dicionário. Para além de erros em português, o ErrorIST 2.0 adapta-se a novos idiomas como o inglês e o francês.

Temos dois cenários distintos onde podemos aplicar futuramente o ErrorIST 2.0 regularmente: a avaliação com alunos universitários e avaliação com a empresa Unbabel. No primeiro cenário, os alunos de Produção de Português Escrito têm como exercício regular corrigir erros em textos e justificar o porquê destes estarem incorretos. No segundo cenário, a Unbabel necessita de avaliar os seus tradutores para garantir aos seus clientes uma tradução fiel. Assim, com a introdução do ErrorIST 2.0 é possível minimizar o tempo e o trabalho investido pelos peritos e, no caso da Unbabel, diminuir, ainda, os custos associados ao longo de toda a avaliação.

1.4 Visão Geral do Documento

Este documento descreve a pesquisa e o trabalho desenvolvido e encontra-se organizado da seguinte forma:

– **Capítulo 2** explicita a teoria que irá suportar o trabalho desenvolvido, como as técnicas utilizadas mais relevantes, e, ainda, a descrição do trabalho anteriormente realizado;

– **Capítulo 3** descreve a primeira versão ErrorIST 1.0 bem como todos os componentes desenvolvidos e necessários ao seu funcionamento;

– **Capítulo 4** apresenta a versão ErrorIST 2.0 e descreve a arquitetura do sistema, detalhando cada um dos componentes, descrevendo a sua implementação e as tecnologias utilizadas;

– **Capítulo 5** mostra o processo de avaliação e os respetivos resultados;

– **Capítulo 6** conclui o trabalho realizado, apresentando os objetivos alcançados e sugestões para trabalho futuro.

Capítulo 2

Trabalho Relacionado

2.1 Taxonomias de Erros

Na literatura existem diversas taxonomias que permitem classificar os diferentes tipos de erros. Segundo Medeiros (Medeiros, 1995), os erros podem ser classificados como cognitivos e tipográficos. Os erros de origem cognitiva são cometidos pela falta de formação linguística do falante ou, muitas vezes, devido à semelhança fonética das palavras. Os erros tipográficos são motivados pela capacidade motora do falante. Estes erros podem ser causados pelo uso do teclado pois, ao pressionar a tecla errada, vai ser produzida uma sequência diferente da esperada.

Ao longo dos anos foram apresentadas diferentes taxonomias de erros como as taxonomias dedicadas a erros humanos, apresentadas por Rasmussen (Rasmussen, 1982) e por Pereira (Pereira, 1983). No entanto, a avaliação em diferentes cenários é um dos objetivos principais do nosso trabalho, o que nos leva a destacar três taxonomias mais relevantes. Nesta Subsecção começamos por rever uma taxonomia básica. De seguida, apresentamos as taxonomias mais direcionadas para a tradução como a *Multidimensional Quality Metrics (MQM)* (Lommel et al., 2014) e a taxonomia do L2F (Costa et al., 2015).

Os tipos de erros apresentados nas taxonomias seguintes encontram-se em anexo, na Secção A.

2.1.1 Taxonomia Básica

A taxonomia mais simples a ter em conta encontra-se na base de todas as outras e apresenta apenas três categorias:

Addition - é adicionada uma letra ou palavra ao texto;

Omission - é removida uma letra ou palavra do texto;

Substitution - é substituída uma letra ou palavra no texto.

A ferramenta GenERRate (Foster, 2009), que será explicada detalhadamente na Subsecção 2.2, fundamenta o seu comportamento nas três categorias apresentadas. Contudo, uma quarta categoria é apresentada, a categoria ***Move***, que se resume na alteração da posição de uma palavra na frase.

2.1.2 Taxonomia *Multidimensional Quality Metrics*

No âmbito da tradução, a taxonomia *Multidimensional Quality Metrics* (MQM) (Lommel et al., 2014) utiliza métricas de avaliação de qualidade para classificar tradutores. A taxonomia mencionada não se centra apenas em erros linguísticos, dá também importância à apresentação do texto.

A taxonomia MQM encontra-se dividida em dimensões[1]: *Design; Locale Convention; Verity; Internationalization; Accuracy; Style; Fluency* e *Terminology*.

A dimensão ***Design*** engloba erros de apresentação do texto, como erros de formatação e estilo de letra. A dimensão ***Locale Convention*** diz respeito a questões relacionadas com formatação de datas, número de telefone e moeda, seguindo as normas a seguir de cada região. A dimensão ***Verity*** refere-se a problemas extra-linguísticos em textos. Um dos exemplos a ter em conta é a tradução de contratos de trabalho, em que a legislação difere de país para país e, nestes casos, poderá ser necessária a intervenção de um profissional com experiência em ambos os sistemas legais para fazer o ajuste de acordo com o país. A dimensão ***Internationalization*** apresenta problemas ligados à internacionalização, como erros de codificação ou produtos indisponíveis em determinados países.

Uma vez que as dimensões anteriormente apresentadas não se relacionam com a proposta de trabalho desenvolvido, apenas as seguintes serão explicadas com mais detalhe. Todos os conceitos apresentados, encontram-se descritos em anexo, na Secção A.

Accuracy

A dimensão *Accuracy*, para além das categorias ***Addition*** e ***Omission***, apresentadas pela taxonomia básica, subdivide-se em *Untranslated* e *Mistranslation*. Os ***Unstranslated errors*** caracterizam-se por apresentar conteúdo no texto traduzido que não foi traduzido, como por exemplo, uma frase em japonês que ao ser traduzida para inglês manteve-se inalterada. Os ***Mistranslation errors*** podem ainda ser especificados em:

- **Ambiguous translation**: O texto traduzido contém ambiguidades resultantes da tradução, fazendo com que o texto tenha uma dupla interpretação;
- **Date/time**: A correspondência de datas ou horários não é consistente no texto original e no texto traduzido;
- **Entity**: Os nomes das entidades mencionadas no texto original não correspondem aos do texto traduzido;
- **False friend**: Uma palavra do texto original é traduzida para uma palavra semelhante, mas errada, no texto traduzido;
- **Technical relationship**: Tradução incorreta de relacionamentos em conceitos técnicos;
- **Number**: Os números do texto traduzido não correspondem aos do texto original;
- **Overly literal**: A tradução é feita de modo excessivamente literal;

[1] http://www.qt21.eu/mqm-definition/definition-2015-12-30.html, último acesso dia 14 Dezembro 2017

- **Should not have been translated**: O conteúdo no texto original não deveria ser traduzido;
- **Unit conversion**: Unidades de conversão e de medida do texto traduzido não correspondem às do texto original.

Style

A dimensão *Style* não está relacionada com erros que alteram o significado do texto, mas sim com erros relacionados com o estilo do texto. Esta dimensão subdivide-se em: *Awkward; Company Style; Inconsistent style; Register; Third-Party style* e *Undiomatic*.

Os ***Awkward errors*** ocorrem quando o texto é afetado por fatores, como o uso excessivo de cláusulas nos contratos de trabalho, que dificultam a sua leitura e compreensão. Os ***Company Style errors*** acontecem quando o texto não está de acordo com as diretrizes impostas pela empresa. Os ***Inconsistent Style errors*** são causados pela inconsistência no texto, quando, por exemplo, um texto é traduzido por mais do que uma pessoa. Os ***Register errors*** correspondem à falta de formalidade (formal ou informal) exigida ou à utilização de gírias inadequadas. Os ***Third-party style*** ocorrem quando um texto não segue as especificações estipuladas como, por exemplo, não seguem um guia linguístico. Um texto traduzido contém ***Undiomatic errors*** quando apresenta traduções gramaticais mas não idiomáticas.

Fluency

A dimensão *Fluency* está ligada a problemas relacionados com a forma ou conteúdo de um textos. De todos os erros apresentados por esta dimensão, apenas iremos destacar os erros que serão fundamentais para a resolução dos requisitos que o sistema proposto visa resolver, como a melhoria dos erros apresentados pelo ErrorIST 1.0 (*Spelling* e *Capitalization*) e a introdução de novos erros ao nível da letra (*Diacritics*) e do discurso (*Cohesion, Inconsistency, Grammatical-Register*).

Salientamos, então, os seguintes erros: ***Spelling, Capitalization, Diacritics, Cohesion, Inconsistency, Grammatical-Register*** e ***Typography***. Destacam-se os ***Spelling errors*** como erros que ocorrem ao nível da letra. Os ***Spelling errors*** são causados pela troca, remoção e adição de um ou mais caracteres numa palavra. Estes erros subdividem-se em ***Capitalization errors*** e em ***Diacritics errors***. Os ***Capitalization errors***, também conhecidos como *truecasing* (Liţă, L and Ittycheriah, A and Roukos, S and Kambhatla, N, 2003), ocorrem quando uma palavra é grafada com letra maiúscula indevidamente e vice-versa e os ***Diacritics errors*** caracterizam-se pela acentuação incorreta de um caracter.

Ao nível do discurso, os erros destacados são os ***Cohesion, Inconsistency*** e os ***Grammatical-Register***. Por vezes, as partes de um texto não estão ligadas entre si ou estão mas de forma incorreta, estes erros são nomeados de ***Cohesion errors***. Os ***Inconsistency errors*** dão-se, ao longo do texto, quando uma entidade é referenciada com expressões diferentes no texto. Os ***Grammatical-Register*** tratam problemas de formalidade na linguagem. Por fim, os ***Typography errors*** relacionam-se com problemas na apresentação do texto. Estes incluem os ***Punctuation errors*** que são caracterizados pelo uso incorreto de pontuação.

Terminology

A dimensão *Terminology* contém erros relacionados com a inconsistência na utilização de uma terminologia específica de domínio ou léxico específico de um cliente, possibilitando uma avaliação mais minuciosa do avaliado.

A dimensão mencionada apresenta erros como: ***Inconsistent with termbase errors, Inconsistent with domain errors, Inconsistent use of terminology errors***. Os ***Inconsistent with termbase errors***, que incluem também os ***Company Terminology*** e ***Third-Party errors***, acontecem quando o texto viola as diretrizes de terminologia da empresa ou de terceiros. Os ***Inconsistent with domain errors*** destacam-se quando um determinado termo não está de acordo com as expetativas do domínio geral. Para finalizar, os ***Inconsistent use of terminology errors*** estendem-se aos ***Multiple terms of concept in source*** e aos ***Multiple translations for the same term errors*** e são criados quando um termo é referido no texto de formas diferentes ou quando é traduzido de forma inconsistente ao longo do texto.

A taxonomia MQM, por ser uma taxonomia abrangente na área de tradução, é utilizada pela empresa Unbabel no processo de categorização dos erros.

2.1.3 Taxonomia do L2F

A taxonomia do L2F (Costa et al., 2015), utilizada no desenvolvimento do sistema ErrorIST 1.0 (dos Santos, 2016), baseia-se em taxonomias criadas por Vilar (Vilar et al., 2006) e Bojar (Bojar, 2011). Vilar dividiu os erros em cinco classes: ***Missing Words, Word Order, Incorrect Words, Unknown Words*** e ***Punctuation Errors***. Mais tarde Bojar, inspirado na taxonomia de Vilar, classificou os erros em quatro tipos: ***"Bad Punctuation", "Missing Word", "Word Order"*** e ***"Incorrect Words"***. Manteve a taxonomia de Vilar, mas eliminou a classe ***"Unknown Words"***. Com base nas taxonomias anteriormente apresentadas, com especial atenção aos erros de tradução e de outros idiomas não cobertos pelas taxonomias anteriores, a taxonomia do L2F classificou os erros em cinco categorias principais: *Orthography, Lexis, Grammar, Semantic* e *Discourse*.

Os *Orthography errors* estão subdivididos em ***Punctuation, Capitalization*** e ***Spelling***, sendo este último pertencente à taxonomia básica. Na taxonomia MQM, os *Orthography errors* encontram-se inseridos na dimensão *Fluency*.

Contrariamente aos ***Orthography errors*** que ocorrem ao nível da letra, os ***Lexical errors*** ocorrem ao nível da palavra. Estes dividem-se em ***Content Words***, ***Function Words*** e ***Unstranslated***. Na taxonomia MQM, apresentada anteriormente, estes erros são categorizados pela dimensão *Accuracy*. Os ***Content Words*** e os ***Function Words errors*** dividem-se em ***Omission*** e ***Addition***. Os ***Content Words*** caracterizam-se pelas palavras que carregam informação na frase, como nomes e adjetivos, enquanto que os ***Function Words*** expressam relações entre palavras, como é o caso dos pronomes ou proposições.

Os aspetos morfológicos e sintáticos de um idioma são categorizados pelos ***Grammar errors*** que,

por sua vez, subdividem-se em ***Misselection*** e ***Misordering***. Tanto os ***Misselection errors*** como os ***Misordering errors*** têm em conta a classe em que cada palavra está inserida. Os ***Misselection errors*** estão, ainda, divididos em ***Verbs, Agreement*** e ***Contractions***, que se relacionam com a conjunção dos verbos e com a concordância das palavras. As trocas entre palavras na frase são classificadas como ***Misordering errors***.

Por vezes, a seleção incorreta das palavras geram problemas ao nível do significado das frases. Assim, os ***Semantic errors*** dividem-se em ***Confusion of Senses, Wrong Choice, Collocational*** e ***Idioms***. Ao traduzirmos a palavra inglesa *"glasses"* para português temos duas traduções possíveis: *"óculos"* e *"copos"*. Se a frase no contexto original se referir a *"copos"* e na frase traduzida a palavra selecionada for *"óculos"*, estamos perante um ***Confusion of Senses error***. Sem confundir com os ***Confusion of Senses errors***, os ***Wrong Choice errors*** distinguem-se pela escolha errada de uma palavra que foi traduzida e que não tem qualquer relação aparente com a palavra original. Os ***Collocational errors*** ocorrem quando numa sequência de palavras, que em conjunto tem um significado, é traduzida cada palavra de modo literal, como por exemplo, na sequência *'get up' (levanta-te)* a tradução literal de cada palavra seria *'obter para cima'*, alterando o significado do texto.

As expressões idiomáticas são muito próprias de cada língua e a tradução de uma expressão à letra, leva a obter uma frase sem sentido. Estes tipos de erros são chamados de ***Idiom errors***.

Para terminar, a taxonomia do L2F apresenta ainda os ***Discourse errors*** que se ramificam em ***Style, Variety*** e ***Should not be translated***. Nos ***Style errors***, um dos exemplos mais concretos é a repetição de palavras. As traduções de um idioma para português por vezes geram ***Variety errors***. Estes erros são criados devido às variações dialetais dda língua como, por exemplo, as variações entre **P**ortuguês **E**uropeu e **P**ortuguês do **B**rasil. Ao traduzir de inglês para português, as estruturas lexicais ou gramaticais poderão ser as do brasileiro. Este erro ocorre frequentemente na tradução de contrações de preposições e artigos (PE: *"Vou **à** praia."* e PB: *"Vou **na** praia."*). Existem expressões e nomes que não devem ser traduzidos de um idioma para outro, como é o caso de nomes ou palavras consideradas como estrangeirismos. Estes erros são apelidados de ***Should not be translated***.

Embora, o ErrorIST 1.0 tenha sido baseado, em parte, na taxonomia do L2F para a classificação dos vários tipos de erros, na nova versão, ErrorIST 2.0, serão disponibilizados mais erros para além destes.

2.2 Software dedicado à geração de erros

A maioria dos sistemas de geração desenvolvidos utiliza a geração de erros para criar dados de treino para serem utilizados em correção automática.

Felice e Yuan (Felice and Yuan, 2014) utilizaram a geração de erros para corrigir erros dados por aprendentes de Inglês como segunda língua. Com base em *corpus* anotados com erros reais, injetaram erros de forma probabilística para criar corpora para corrigir os vários tipos de erros. De modo a refinarem os contextos onde ocorrem os erros e a inserirem os erros com mais precisão, utilizaram um

Part-of-Speech (POS) *tagger*, que atribui etiquetas morfológicas ao texto. Por fim, os autores concluiram que, apesar dos resultados variarem de acordo com os vários conjuntos de treino, a geração de erros baseada em erros reais e a utilização de um POS *tagger* melhora o desempenho dos sistemas de correção. No entanto, encontram-se na literatura alguns sistemas dedicados à geração de erros como o GenERRate (Foster, 2009) e o Missplel (Bigert et al., 2003), que dado um texto são capazes de gerar e inserir vários tipos de erros nesse mesmo texto.

2.2.1 GenERRate

O GenERRate (Foster, 2009) é uma ferramenta de geração de erros que transforma uma frase bem formada numa frase com erros. Esta ferramenta baseia-se na taxonomia básica, descrita na Subsecção 2.1.1, para gerar os erros. Assim, a geração dos erros é feita através de quatro tipos de operações genéricas: inserir, apagar, substituir e mover.

A inserção dos erros pode ser feita aleatoriamente, com base numa lista de palavras e ainda através da identificação das etiquetas morfológicas atribuídas por um *POS tagger*.

- **Erros inseridos aleatoriamente**
  É selecionada, aleatoriamente, uma palavra numa frase e esta, de acordo com as diferentes operações, pode ser removida, substituída ou movida para outra posição.

- **Erros inseridos aleatoriamente com base numa lista de palavras**
  Recorrendo a uma lista de palavras, podemos inserir uma nova palavra, dessa lista, numa posição aleatória. Esta lista é também utilizada para operações de substituição em que a palavra, selecionada aleatoriamente, é substituída por outra da lista.

- **Erros inseridos com base no POS tagger**
  Através de um *POS tagger* é possível inserir ou substituir uma palavra com a mesma etiqueta ou até mesmo com uma etiqueta diferente. O mesmo acontece se quisermos apagar e mover uma palavra com uma determinada etiqueta. Com base nas etiquetas das palavras vizinhas, o utilizador pode especificar que tipo de etiqueta pretende para gerar o erro.

2.2.2 Missplel

O Missplel (Bigert et al., 2003) é uma ferramenta de geração de erros que se encontra dividida em quatro módulos principais.

Tal como o GenERRate, os erros gerados pelo Missplel são erros que seguem a taxonomia básica, pois baseiam-se em operações como inserir, remover, substituir e mover para gerarem erros ao nível da letra e da palavra.

O primeiro módulo do Missplel, ***Damerau***, introduz erros de *Damerau* (Damerau, 1964) conhecidos como ***Spelling errors***. Estes erros são provocados pelo uso do teclado, por exemplo *'escola'* e *'escoal'*, em que existe uma troca da letra *'l'* com a letra *'a'*. O utilizador pode controlar o fluxo de erros deste tipo, criando outras palavras da mesma forma.

O segundo módulo, ***Split Compound***, dedica-se a criar erros relacionados com palavras compostas. Um exemplo deste tipo de erros é a transformação da palavra *'blackboard'* para *'black board'*.

O terceiro módulo é o ***Sound Error***. Insere erros ao nível da letra e permite que erros de competência[2], ou seja, erros cometidos por não se conhecer a forma de escrita correta, sejam facilmente introduzidos. Um exemplo deste tipo de erro é transformar *'bee'(abelha)* em *'be' (verbo ser)* .

O quarto módulo, o ***Syntax***, faz a inserção dos erros ao nível da letra ou da palavra. Pode ser usado para introduzir erros de ordem de palavras, erros de concordância, erros de tempo verbal, erros de repetição ou erros de omissão. Infelizmente, não é possível explicar em pormenor os módulos apresentados, pois a informação encontrada não está suficientemente detalhada.

Neste sistema, cada erro tem uma probabilidade associada. Assim, permite que erros comuns de ortografia sejam introduzidos com mais frequência. Todos os erros gerados pelo Missplel são gerados aleatoriamente, tirando assim alguma liberdade de escolha ao utilizador quanto ao tipo de erro.

2.3 Software dedicado à correção de erros

Nesta Sub-Secção, apresentamos alguns corretores de erros como o Correcto, o *Language Tool*, o GNU *Aspell* e o *Grammarly*.

Para textos de língua portuguesa, destacamos o Correcto (Medeiros, 1995), um corretor ortográfico que utiliza um analisador morfológico, o Palavroso. A utilização deste analisador morfológico facilita o processo de alteração de radicais dos verbos através das regras de transformação aplicadas pelo mesmo. Medeiros, classificou as regras de transformação em dois tipos distintos: regras de conversões regulares, relacionadas com regras de ortografia, e regras de formas verbais, que são utilizadas para analisar formas de verbos irregulares com base em regras dos verbos regulares. Uma das inovações deste trabalho foi o tratamento de erros que envolvem verbos com enclíticos, onde o verbo e um pronome se juntam através de um hífen (enclítico).

Os corretores de erros como o *Language Tool*[3] e o GNU Aspell[4] encontram-se disponíveis para várias línguas. Para além de corretor, o *Language Tool*, contém regras gramaticais que abrangem mais de 25 idiomas como o francês e o russo. São ainda disponibilizados alguns conjuntos de erros mais frequentes em inglês.

O GNU *Aspell* consiste numa versão melhorada do *Ispell*[5] e tem como principal característica sugerir possíveis correções. Este corretor ortográfico livre também pode ser utilizado como bilbioteca e suporta a maioria dos idiomas ocidentais. Contrariamente ao pioneiro *Ispell* o *Aspell* consegue processar textos com formatação UTF-8[6] sem recorrer a um dicionário especial, suporta vários dicionários em simultâneo e manipula dicionários pessoais. Apesar de ser uma versão melhorada do *Ispell*, a interface foi mantida de modo a que seja possível integrar com o mesmo conjunto de aplicações.

[2]https://motivatedgrammar.wordpress.com/tag/competence/, último acesso dia 14 Dezembro 2017
[3]http://wiki.languagetool.org, último acesso 10 de Maio 2018
[4]http://aspell.net, último acesso 25 de Julho 2018
[5]https://www.gnu.org/software/ispell/, último acesso 10 Dezembro 2017
[6]https://www.w3schools.com/charsets/ref_html_utf8.asp

O *Grammarly*[7] é também um corretor de texto automático baseado em regras e é utilizado apenas para o inglês. Deteta erros de vários tipos e apresenta sugestões de correção ao utilizador. Para além da correção, o *Grammarly*, atribui a cada utilizador uma pontuação baseada no número de erros, para que o utilizador obtenha uma avaliação da sua escrita. Com o *Grammarly Handbook* é possível vizualizar todas as regras aplicadas à língua inglesa servindo de gramática de apoio aos utilizadores.

A análise destes corretores ser-nos-à útil para a criação de um mecanismo capaz de gerar erros através da observação da aplicação das regras de correção de forma inversa.

2.4 Outros Recursos

O COPLE2 – *Portuguese Learner Corpus* (Mendes et al., 2016), um *corpus* de apoio à geração de erros comuns da língua portuguesa, inclui textos corrigidos que foram escritos e falados por aprendentes que utilizam o português como segunda língua ou língua estrangeira. Os textos estão transcritos para um formato digital que contém as correções feitas pelo avaliador. Como exemplo, temos a seguinte frase retirada do COPLE2, onde podemos visualizar o tipo de erros cometidos:
"Eu já ****vivi**** *(vivo) em Portugal* ****para**** *(há) um* ****año**** *(ano)!"*.

2.5 Discussão

A inserção de erros automáticos já tinha sido implementada por sistemas descritos na Subsecção 2.2, mas com o objetivo de criar corpora para aprendizagem automática. O ErrorIST difere por ser um sistema que se destina a avaliação linguística.

As ferramentas como o GeneRRate e o Missplel introduzem tipos de erros mais genéricos, seguindo as operações apresentadas pela taxonomia básica, descrita na Subsecção 2.1.1. A versão anterior, ErrorIST 1.0, apresentava tipos de erros da taxonomia do L2F orientados à tradução automática. Nesta nova versão, o ErrorIST 2.0 disponibilizará um cardápio de erros baseados não só na taxonomia do L2F, mas também noutras taxonomias de erros como a MQM. As regras dos sistemas de correção como o *Language Tool* e o *Grammarly* e ser-nos-ão úteis na medida em que ao invertermos as regras de correção, aplicadas no sistema, obtemos regras de geração para erros. Sendo o *Language Tool* um corretor disponível em mais de 25 línguas, este irá ser utilizado para regras do português e inglês. O *Grammarly Handbook* será utilizado para desenvolver regras para o inglês, pois é a única língua suportada pelo sistema. Em cada avaliação do ErrorIST 2.0 serão fornecidos, por peritos, dados reais de erros frequentes numa dada língua, que servirão de apoio ao desenvolvimento de novos tipos de erros.

[7] https://www.grammarly.com/faq#toc0, último acesso 13 Maio 2018

Capítulo 3

ErrorIST 1.0

O ErrorIST 1.0 (dos Santos, 2016) é uma ferramenta, desenvolvida em Python, que tem como objetivo avaliar falantes através da geração de erros em textos. Dado um texto e os tipos de erro a inserir, o ErrorIST 1.0 gera os erros e adiciona-os ao texto. Depois de introduzidos os erros nos textos, estes são submetidos a falantes para correção. Por fim, o ErrorIST 1.0 avalia o falante consoante o seu desempenho.

3.1 Arquitetura

O ErrorIST 1.0 encontra-se dividido em três módulos principais: *Error Generator, Tracer* e *Evaluator*, como apresentado na Figura 3.1.

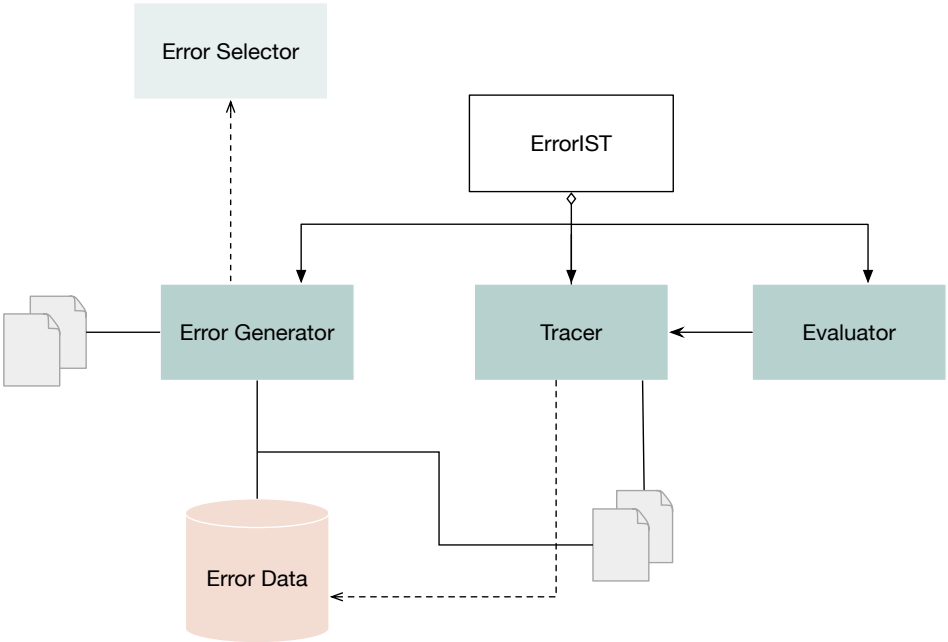

Figura 3.1: Arquitetura do ErrorIST 1.0

3.1.1 Error Generator

O *Error Generator* utiliza um subconjunto da taxonomia do L2F, descrita na Subsecção 2.1.3 para gerar os vários tipos de erros. No *Error Selector*, cada utilizador deve especificar o tipo e o número de erros para que o *Error Generator* insira os erros de acordo com as especificações pretendidas. Ao receber um texto como *input*, o *Error Generator* gera os erros, adicionando-os posteriormente ao texto. É inserido um erro por frase e sempre que este não se aplique à frase o gerador passa à frase seguinte. Este módulo recebe ainda como *input* ficheiros de recurso, descritos na Subsecção 3.2.

Os erros recebidos pelo *Error Selector* podem ser erros ortográficos, lexicais, gramaticais, e semânticos. De seguida serão apresentados os erros, baseados na taxonomia do L2F, disponibilizados pelo ErrorIST 1.0.

- **Orthography Errors**

  Os *Orthography errors*, como já foi apresentado na taxonomia do L2F, na Subsecção 2.1.3, dividem-se em erros de ***Punctuation, Capitalization*** e ***Spelling***. Relativamente aos ***Punctuation errors***, o ErrorIST 1.0 adiciona e remove vírgulas, ponto e vírgulas e pontos. De modo que os erros de pontuação se assemelhem mais aos erros reais, é necessário utilizar o ficheiro POS *tagger* que irá identificar as classes de palavras de cada frase. Depois de ser feita a identificação, irá inserir vírgulas, ou ponto e vírgula, entre um sujeito e um predicado ou entre um verbo e os seus complementos. Caso a seguir a um verbo exista um adjetivo, substantivo, pronome ou artigo, o ErrorIST 1.0 insere uma vírgula após o verbo. As vírgulas também poderão ser removidas nos casos em que ocorrem depois de um advérbio, ou seja, o *POS tagger* identifica um advérbio e o caractere que se segue, se este for uma vírgula, será removido. Os ***Capitalization errors*** apenas alteram a primeira letra de uma palavra de minúscula para maiúscula e vice-versa. Nos ***Spelling Errors*** é possível omitir, adicionar, substituir ou alterar a ordem das letras. Alguns dos erros mais comuns são erros ortográficos como, por exemplo, trocar *'z'* por *'s'*. Para criar mais erros deste tipo o utilizador pode acrescentar mais regras ao ficheiro ***Sounds File***.

- **Lexical Errors**

  Para criar um ***Omission error***, o ErrorIST 1.0 recorre ao POS *tagger* e, através da etiqueta morfológica, remove a palavra. Este processo é efetuado tanto para ***Omission Function Words*** (ex.: conjuções) como para ***Omission Content Words*** (ex.: verbos e nomes). Como referido anteriormente, o ficheiro *Word File* contém uma lista de palavras que está organizada por etiquetas, que será apresentado em detalhe na Subsecção 3.2. Para introduzir um ***Addition error***, o ErrorIST 1.0 escolhe uma palavra da lista e insere-a na frase. Ambos os processos de seleção e de inserção da palavra são feitos aleatoriamente. Considerando a frase *"Eu corri a maratona."*, se inserirmos um ***Addition Function Error***, o ErrorIST 1.0 seleciona uma das etiquetas (ex.:

conjunção), e insere-a na frase *"Eu ***quando*** corri a maratona."*.

– **Grammar Errors**

Os ***Grammar Errors*** estão divididos em dois grupos principais: ***Misselection*** e ***Misordering***. Os ***Misselection Errors*** utilizam a informação contida nos *Affix Files* para alterarem as palavras através de afixos. Em relação aos ***Verb Tense Errors*** e aos ***Verb Person Errors***, o ErrorIST 1.0 utiliza a informação dos *Affix files* e altera o sufixo, mantendo o tempo e a pessoa, consoante a escolha do utilizador. Os ***Verb Blend Errors*** modificam os sufixos, mas sem a preocupação de manter o tempo ou a pessoa. Devido à limitação do uso dos *Affix files*, o sistema, na maior parte das vezes, é incapaz de garantir a inserção dos verbos irregulares corretamente. Considerando o exemplo em que temos um verbo irregular, o verbo *"ir"*, ao aplicarmos as seguintes regras do *Affix File*:

```
I R >-R, STE # \P=2, N=s, T=pp"
```

O resultado será *"iste"*, e esta é uma palavra inexistente no dicionário de língua portuguesa. Por outro lado, este tipo de erros poderá ser proveitoso em dados contextos, como por exemplo na avaliação com crianças ou estrangeiros. Os *Affix Files* contêm categorias que são utilizadas para os ***Word Class error***, onde a cada categoria é atribuída uma etiqueta. Desta forma, depois de analisar a palavra, é aplicada uma regra ao seu sufixo que irá alterar a sua categoria. Exemplo de algumas regras do *Affix file*:

```
Flag n ; \CAT=v, T=inf"
```

```
I R >-IR, ENTE ; \CAT=adj, N=s, FSEM=nte"
```

```
O R >-IR, ENTE ; \CAT=adj, N=s, FSEM=nte"
```

Estas transformações podem ser aplicadas a verbos (`CAT=v`) que através da aplicação das regras a categoria será alterada para adjetivo (`CAT=adj`). Por exemplo, o verbo *"aderir"*, ao aplicar a primeira regra, será transformado no adjetivo *"aderente"*.

O ErrorIST 1.0, para a inserção dos ***Agreement errors***, recorre aos *Affix Files: Number, Gender* e *Person*. Na inserção de ***Agreement Gender***, se a palavra estiver no plural, antes da alteração do género, é necessário transformá-la em singular e, no fim, voltar a colocá-la no plural. O número de palavras existente para a geração dos ***Agreement Person Errors*** é muito pequeno, pelo que não se justifica a utilização de um ficheiro do tipo *Affix File*. No entanto, o utilizador deverá fornecer o ficheiro com as palavras pretendidas e com o formato apresentado na Subsecção 3.2. À semelhança dos ***Agreement Person errors***, os ***Contractions errors*** são inseridos com base no *Contraction File* que contém a contração e a respetiva decomposição, como está apresentado na Subsecção 3.2. Os ***Misordering errors*** são trocas entre as palavras na frase, ou seja, são escolhidas duas palavras, aleatoriamente, e estas trocam entre si de modo a alterarem a frase.

– **Semantic Errors**

Como este projeto se encontra mais focado para a tradução, a inserção de ***Confusion of Sense errors*** torna-se mais complexa. É preciso ter em conta o idioma de origem. Aqui, o uso do *Wiktionary*[1] é fundamental, pois contém uma página para cada palavra com as respetivas traduções. Para introduzir um ***Confusion of Sense Error***, o ErrorIST 1.0, recorre ao *Wiktionary* para ver as traduções associadas à palavra e selecciona uma. De seguida, faz o mesmo para a palavra seleccionada. Temos como exemplo a palavra *"rede"*, na frase *"Rede Social"*, onde a tradução para inglês é *"Social network"*, poderá ser traduzida para *"armadilha"*, ficando a frase *"Armadilha social"*.

3.1.2 Tracer

Após o *Error Generator* introduzir os erros no texto é submetido a uma correção feita pelo falante. O *Tracer* recebe como *input* um ficheiro com a correção feita pelo avaliado e identifica as intervenções feitas ao ficheiro e avalia se:

– O editor modificou o erro inserido (`Modificado`);

– O editor modificou o erro para a forma original (`Esperado`);

– O editor modificou a frase mas não modificou o erro inserido (`Outro`);

– O editor não detetou o erro e não fez nenhuma alteração (`Não Modificado`).

3.1.3 Evaluator

Após a verificação do *Tracer* e a avaliação do estado do erro, como `Modificado`, `Esperado` ou `Outro`, o *Evaluator* avalia a edição do falante como `OK` (correta), `INDEF` (inesperada) e `KO` (incorreta). Considerando a introdução de um ***Spelling Error*** na frase *"A Mariana segura a mala"*, obtemos a frase *"A Mariana segura a ***mlaa***"*. Tendo em conta o exemplo, a avaliação seria:

– `OK` – a frase foi modificada e o resultado era o esperado, nao existindo outras alterações, isto é, o editor corrige *"mlaa"* para *"mala"*;

– `INDEF` – a frase foi modificada, mas não da forma esperada, por exemplo *"A Mariana segura a ***carteira***"* (é necessária uma verificação humana nestes casos);

– `KO` – a frase com o erro manteve-se inalterada.

A avaliação feita pelo *Evaluator* encontra-se resumida na Tabela 3.1.

[1] https://en.wiktionary.org/, último acesso 4 Dezembro 2017

Tabela 3.1: Avaliação feita pelo *Evaluator*

Modificado	Esperado	Outro	**Resultado**
Sim	Sim	Não	OK
Sim	Sim	Sim	INDEF
Sim	Não	Não	INDEF
Sim	Não	Sim	INDEF
Não	-	Sim	INDEF
Não	-	Não	KO

3.2 Recursos utilizados pelo ErrorIST 1.0

O ErrorIST 1.0 utiliza recursos externos para gerar erros, como os ***Verb Errors***, os ***Addition Errors***, os ***Confusion of Senses Errors*** e os ***Agreement Errors***. Os recursos utilizados pelo ErrorIST 1.0 são fundamentais para a identificação das palavras no texto e serão descritos, de seguida, com maior detalhe, evidenciando a influência de cada um no funcionamento da ferramenta descrita.

3.2.1 POS *tagger*

O *Part-of-Speech tagger* ou POS *tagger* faz a atribuição de uma etiqueta morfológica às palavras consoante a sua função na frase. O POS *tagger* utilizado pelo ErrorIST 1.0 deve seguir as mesmas diretrizes que os apresentados na biblioteca NLTK (Bird, 2006) e deve ser fornecido pelo utilizador ao ErrorIST 1.0, consoante o idioma pretendido. A maioria dos tipos de erros implementados pelo ErrorIST 1.0 são gerados com base neste processo de etiquetagem das palavras. Dada a frase *"O Ivo comprou um telemóvel."*, um exemplo da identificação feita pelo POS *tagger* seria:

```
[('O', 'art'), ('Ivo', 'n'), ('comprou', 'v-fin'), ('um', 'art'), ('telemóvel', 'n')]
```

3.2.2 Word File

Para a introdução de erros do tipo ***Addition Errors*** é necessário recorrer ao *Word File*, um ficheiro composto por várias palavras, disponibilizado pelo utilizador. Cada linha do ficheiro contém a etiqueta morfológica seguida das palavras associadas a essa etiqueta. O ficheiro deverá seguir o seguinte formato:

```
ART a o uma
NOUN cão coelho mota colher
VERB rir comer sorrir
```

Ao inserir o erro, o ErrorIST 1.0 recorre a esta lista de palavras e escolhe uma palavra aleatoriamente. Se o avaliador pretender inserir um ***Omission Error*** com a etiqueta `'art'` (artigo), é removida da frase uma das palavras correspondentes à etiqueta especificada.

3.2.3 Affix Files

Os *Affix files*, ou ficheiros de afixos, contêm regras provenientes do *Ispell* compostas por expressões regulares. Uma expressão `'regex'` contém a sequência a remover `'-removal'`, a nova sequência para substituir `'substitution'` e a informação da alteração do sufixo `'info a=x, info b=y'`, tal como no exemplo seguinte:

```
regex >-removal, substitution; ''info a=x, info b=y''
```

Para cada regra, é criada uma entrada reversível que permite que o sufixo seja adicionado ou removido, garantindo que a palavra volte novamente ao estado original, como por exemplo, para que um verbo volte a estar no infinitivo. A sequência seguinte mostra um exemplo de uma entrada reversível:

```
I R >-R, AM # ''P = 3, N = p, T = pi, Reverse = yes''
```

Com a aplicação desta regra, podemos alterar o tempo dos verbos que terminam em `'I R'`, retirando a letra `'R'` e adicionando as letras `'AM'`. Os parâmetros que se seguem ao `'#'` indicam a pessoa (`'P = 3'`, 3ª pessoa), o número (`'N = p'`, plural) e o tempo verbal (`'T = pi'`, pretérito imperfeito) e se a regra é reversível (`Reverse = yes''`). Aplicando esta regra à verbo *"sorrir"*, obtemos a palavra *"sorriam"*. Caso a palavra incial seja *"sorriam"* poderemos aplicar a regra e obtemos *"sorrir"*.

Os verbos irregulares, devido às suas terminações não seguirem o padrão regular de conjugação, ao aplicar a regra, poderão ser criadas palavras inexistentes.

3.2.4 Sounds File

Os ***Spelling Errors***, para além de troca de letras, omissões, repetições e adições, apresentam erros sonoros. Estes erros caracterizam-se por uma correspondência entre as letras e os sons. Uma letra pode representar mais do que um som. O som da letra *'c'*, por exemplo, pode ser confundida com os sons das letras *'k', 's'*.

O ficheiro *Sounds File* contém sequências de homofonia separadas por um espaço em branco. Se, durante a inserção do erro, estiver presente uma sequência do ficheiro com as sequências disponibilizadas pelo avaliador, o ErrorIST 1.0 substituirá a palavra pela nova sequência:

```
's' 'z'
```

A palavra *"análise"* com a aplicação das regras ficaria *"análize"*. No entanto, podem ser produzidos uma série de erros devido à devido à correspondência não unívoca entre grafemas e fones (Zorzi, 2008) presentes na língua.

3.2.5 Wiktionary

Wiktionary[2] é um dicionário multilingue livre, disponível para cerca de 172 línguas. Contrariamente aos dicionários habituais, este dicionário está inserido num projeto colaborativo, onde qualquer pessoa

[2]https://en.wiktionary.org/, último acesso 4 Dezembro 2017

pode editar e guardar definições. Ao pesquisarmos uma palavra no *Wiktionary*, conseguimos aceder às respetivas traduções em várias línguas e a outras informações da palavra, tais como sinónimos e antónimos. No ErrorIST, este dicionário colaborativo é utilizado para a geração dos ***Confusion of Sense Errors***, apresentados na Subsecção 4.1.2.

3.3 Avaliação

A ferramenta ErrorIST 1.0 foi avaliada, pelo o autor anterior **??**, quanto ao seu desempenho, à qualidade de inserção de erros e à capacidade de avaliação de falantes.

Para testar a qualidade do ErrorIST 1.0, foram gerados 6 erros de cada tipo por texto, segundo os métodos descritos na Subsecção 4.1.2. A avaliação foi feita por um perito que verificou se os 126 erros gerados estavam de acordo com a descrição da taxonomia do L2F. A maioria dos erros gerados estava correta e atingiu os 66.67%. O autor conclui que erros como ***Verb Blend, Agreement: Gender, Blend*** e ***Confusion of Senses***, foram inseridos poucas vezes com sucesso (33.33%). Concluiu, ainda, que um dos motivos para o insucesos destes erros é não haver distinção entre verbos regulares e irregulares na aplicação das regras dos ficheiros de afixo, gerando, assim, palavras que não existem.

O segundo teste foi feito com 10 falantes e foram selecionados dois textos. O primeiro texto consistia numa história escrita em português e o segundo texto continha legendas de filmes traduzidas de inglês para português. Para a avaliação, os falantes foram divididos em grupos de dois e cada grupo corrigiu um dos textos após a geração dos erros. Após a correção, foi atribuída uma pontuação tendo em conta a avaliação feita pelo *Evaluator*, que avalia se o erro foi corrigido da forma esperada ou não, como apresentado na Tabela 3.1. Com os resultados obtidos pelo ErrorIST 1.0 concluiu-se que 66,83% das vezes foi possível classificar as intervenções dos falantes como corretas e incorretas. As restantes intervenções classificadas como indefinidas necessitaram de verificação humana, pois são alterações que foram feitas, mas que não eram esperadas pelo ErrorIST 1.0.

De modo a avaliar a qualidade das intervenções feitas pelos falantes, o ErrorIST 1.0 foi anteriormente testado pela Unbabel, num ambiente empresarial. A amostra de texto continha 17 frases retiradas de e-mails traduzidos, e continham um erro de cada um dos seguintes tipos: ***Punctuation, Agreement: Verb Tense, Confusion of Senses, Capitalization, Contraction*** e ***Spelling***. Do texto, 7 das 17 frases não continham erro, apenas serviram para verificar o número de vezes que os falantes alteravam a frase. Com a utilização do ErrorIST 1.0 a Unbabel concluiu que 58.52% das correções podem ser detetadas automaticamente por este sistema.

Capítulo 4

ErrorIST 2.0

4.1 Arquitetura

Na nova versão do ErrorIST, 2.0, entendemos que seguir a arquitectura existente seria o mais indicado. No entanto, perante a dificuldade encontrada ao verificar a existência das palavras geradas pelo ErrorIST 1.0, foi acrescentado um novo módulo opcional, o módulo *Checker*, de acordo com a Figura 4.1. Este módulo, verifica a existência das palavras geradas, com recurso aos *Affix Files*, pelo *Error Generator*. Após a inserção do erro, o módulo *Tracer* deteta as alterações efetuadas pelo editor que, posteriormente, são avaliadas pelo *Evaluator*.

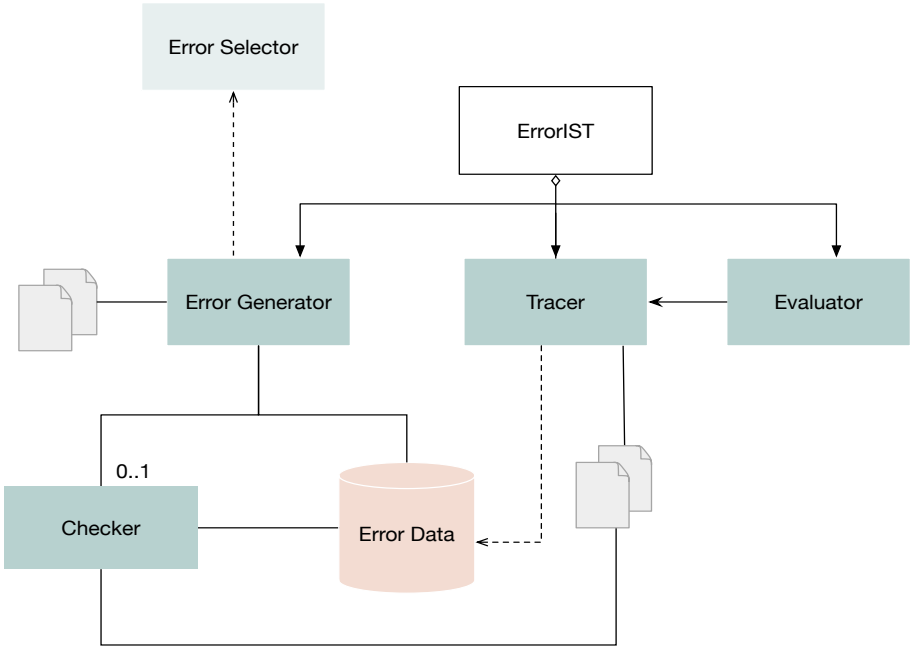

Figura 4.1: Arquitetura do ErrorIST 2.0

4.1.1 Error Selector

Neste módulo, o avaliador insere como *input* os tipos e o número de erros pretendidos que serão gerados pelo ErrorIST 2.0 e introduz os ficheiros, descritos nas Subsecções 3.2 e 4.2, correspondentes a cada idioma, através da linha de comandos da seguinte forma:

```
- <TIPO DE ERRO 1> <NÚMERO DE ERROS> ...  <TIPO DE ERRO N> <NÚMERO DE ERROS>
-SPELLING 6 -VERB TENSE 3
```

Comparativamente à versão anterior do ErrorIST 2.0, foi introduzido o conceito de erros *"à la carte"* onde, apesar do sistema se basear em muitas taxonomias, não é seguida uma taxonomia em específico. Este conceito permite a integração de novos tipos de erros no sistema e permite, ao utilizador, refinar a sua avaliação. Para além dos erros disponibilizados pelo ErrosIST 1.0, descritos na Subsecção 4.1.2, foram introduzidos erros de: utilização do teclado, ***Keyboard*** e ***Whitespace Errors***; acentuação gráfica, ***Graphic Accentuation Errors***; omissão e substituição de preposições, determinantes e pronomes, ***Wrong Preposition/Determiner***, ***Omission Preposition/Determiner/Pronoun***; erros comuns do português, inglês e francês, ***Commons Errors***; próclise, ***Proclise Errors***; variações do português ***Wrong Language Variety***; formalidade e registo no discurso ***Register***.

4.1.2 Error Generator

No módulo *Error Generator* são gerados os erros selecionados no *Error Selector*, juntamente com os ficheiros de recurso, como apresentado na Figura 4.2. Para além dos ficheiros anteriormente recebidos, descritos na Subsecção 3.2, foram adicionados ficheiros necessários à geração dos ***Wrong Preposition*** e ***Wrong Determiner***, ***Common Errors*** e ***Wrong Language Variety***.

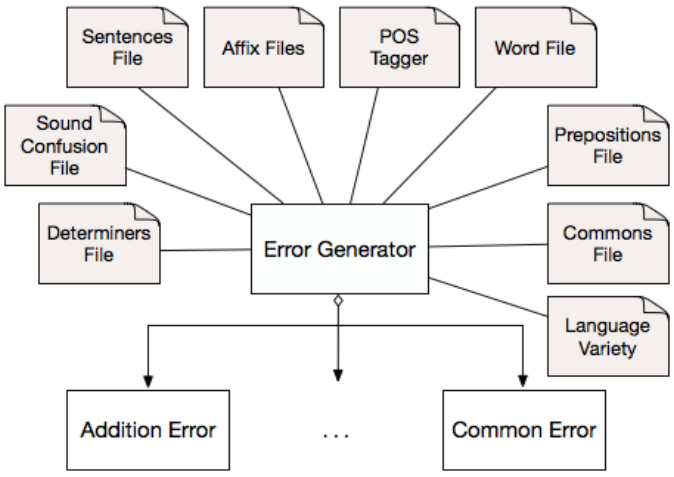

Figura 4.2: Arquitetura do Error Generator

De acordo com a Figura 4.2, o *Error Generator* recebe como *input* os seguintes ficheiros:

- ***Sentence File*** – texto onde vão ser inseridos os erros e que contém uma frase por linha;
- ***POS tagger*** – identificador morfológico de acordo com o idioma;
- ***Word File*** – contém palavras para os *Addition Errors* de acordo com o idioma;
- ***Sound Confusion File*** – contém as sequências de sons que serão utilizadas em erros de *Spelling*;
- ***Affix Files*** – conjunto de ficheiros utilizados na alteração dos afixos dos *Verbs e Agreement Errors* de acordo com o idioma;
- ***Commons File*** – contém palavras correspondentes aos tipos de erros mais frequentes de acordo com idioma;
- ***Prepositions File*** – contém todas as preposições de acordo com o idioma;
- ***Determiners File*** – contém todos os determinantes de acordo com o idioma;
- ***Language Variety File*** – contém palavras correspondestes à variação dos idiomas;

O Error Generator irá gerar os vários tipos de erros, juntamente com os ficheiros de recurso, e inseri-los em cada uma das frases.

Como *output*, o *Error Generator* irá devolver um ficheiro para que o avaliado possa editar, *Correction File*, com todos os erros inseridos e o *Verification Object* para cada um dos erros, que contém as informações sobre o erro introduzido. Estes, serão armazenados no *Error Data* para que, depois das edições dos avaliados, possam ser acedidos pelo *Tracer*.

Caso nenhum tipo de erro se adeque à frase, o *Error Generator* passará à frase seguinte.

Após a realização de vários testes ao ErrorIST 1.0, detetámos alguns aspetos a melhorar no modo de geração de alguns tipos de erros existentes e a necessidade de implementar outros tipos de erros.

4.1.2.1 Melhoria dos erros existentes

Capitalization

Muitas línguas fazem a distinção entre letras minúsculas e maiúsculas, no entanto, o grafar de maiúscula não se aplica a idiomas que não derivem do latim, grego ou cirílico como, por exemplo, o árabe, o japonês e o tailandês. No processo de verificação de gramática, a primeira letra em maiúscula fornece informações importantes para a desambiguação de palavras como o reconhecimento de entidades nomeadas, uma vez que os nomes próprios são grafados com maiúscula. Contudo, existem exceções à regra e, no inglês, o pronome *'I'* é representado por uma letra maiúscula. Outra excepção é no Alemão, onde as letras maiúsculas não são utilizadas apenas em nomes, neste caso o ErrorIST não é capaz de inserir *Capitalization Errors* para além dos nomes.

O problema apresentado pelos ***Capitalization Errors*** era o facto de o grafar de maiúscula poder ser feita, por exemplo, em verbos e determinantes. Para evitar estes casos, é atribuída uma etiqueta morfológica à palavra e a palavra é grafada com maiúscula ou minúscula apenas em nomes e na excepção apresentada pelo inglês.

Punctuation

Na geração dos ***Punctuation Errors*** era possível que caracteres como '@' e '#' fossem inseridos como pontuação. Este tipo de inserção não se adequa ao tipo de escrita habitual de um texto, tanto em português, como em inglês e francês e, desta forma, foram retirados todos os caracteres que não são considerados caracteres de pontuação.

Verbs: Tense, Person e Blend

A geração dos erros de conjugação de verbos é realizada com recurso a *Affix Files*, descritos na Subsecção 3.2 e, por vezes, são geradas palavras inexistentes, chamando assim à atenção do avaliado.

Consideremos a frase seguinte: *"Se eu fosse aí, trazia bolo"*. Com a versão ErrorIST 1.0 é possível gerar a palavra **"*trazi*"**. Para solucionar este problema, a existência da palavra gerada é verificada através do módulo *Checker*, um módulo opcional, que será apresentado na Subsecção 4.1.3. A palavra **"*trazi*"** deixa de ser introduzida como erro e é gerada uma nova palavra até esta seja validada pelo módulo *Checker* como, por exemplo, a palavra **"*traziam*"**.

4.1.2.2 Introdução de novos erros

Keyboard

Os ***Keyboard Errors***, também apresentados por Missplel (Bigert et al., 2003) e por José Carlos Medeiros em (Medeiros, 1995) como erros de Damerau (Damerau, 1964), estão relacionados com a capacidade psico-motora do escritor e com o teclado utilizado. As trocas de caracteres são muito frequentes quando o falante utiliza um teclado para escrever. Apesar de fazerem parte dos ***Spelling Errors***, contrariamente a estes, não se baseiam nos caracteres constituintes de cada palavra ou em caracteres aleatórios. Os ***Keyboard Errors*** são mais específicos e têm em conta a disposição que o teclado de cada idioma apresenta. Embora exista imprevisibilidade neste tipo de erros (Medeiros, 1995; *"Por serem de origem motora, são aleatórios e imprevisíveis.", página 84*), se o teclado for devidamente identificado pelo utilizador é possível prever determinados erros através da troca de letras vizinhas. Na Figura 4.3 temos um exemplo de um teclado QWERTY, onde podemos observar a distância entre as teclas:

Um exemplo de um ***Keyboard Error***, gerado pelo ErrorIST 2.0, na frase *"A Maria fez um bolo de chocolate."* é *"A Maria fez ***im*** bolo de chocolate."* onde a letra *'u'* é trocada pela letra *'i'* à sua direita no teclado.

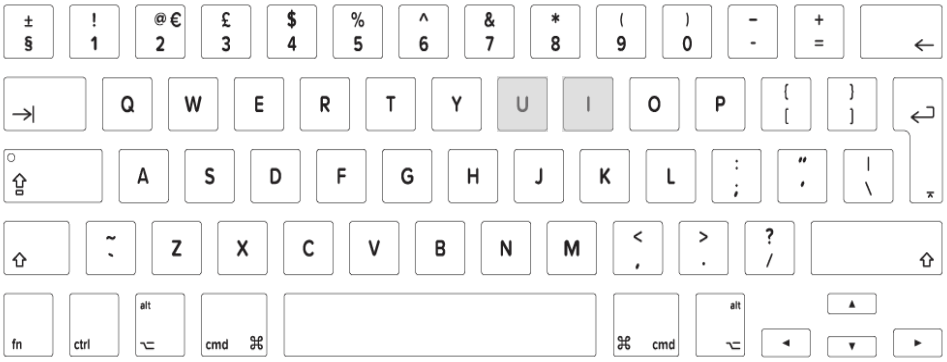

Figura 4.3: Exemplo de um *Keyboard Error* num teclado QWERTY

Um erro comum de teclado, e também introduzido pelo ErrorIST 2.0, é o ***Whitespace Error*** onde é adicionado um espaço, para além do existente, entre as palavras.

Graphic Accentuation

A acentuação é fundamental para representar as regras gramaticais de alguns idiomas. Os *Graphic Accentuation Errors*, também conhecidos como ***Diacritics Errors*** na taxonomia MQM (Lommel et al., 2014), descrita na Subsecção 2.1.2. Este tipo de erros é inserido através da omissão, adição e troca de diacríticos. O seguinte exemplo apresenta um ***Graphic Accentuation Error*** em francês, onde a palavra correta é *'déjà'*, com dois acentos, um agudo e outro grave, e o ErrorIST 2.0, gera a palavra *'*déjá*'*, trocando o último acento por um agudo.

Commons

Em todos idiomas existem erros que ocorrem com mais frequência e que, por esse motivo, se destacam na avaliação da escrita de textos. Com recurso aos corretores automáticos *Grammarly* e *Language Tool*, descritos na Subsecção 2.3, que contêm regras de correção, foram criadas regras que, inversamente, nos permitem gerar erros. O *Grammarly Handbook* tem regras para o inglês necessárias para introduzir alguns tipos de erros como, por exemplo, erros de contrações como *'it's'*, que o ErrorIST 2.0 modifica para *'*its*'*.

Os ***Common Error*** são gerados com recurso a ficheiros que contêm sequências de modificações de erros. Se a sequência estiver presente na frase durante a inserção de um ***Common Error***, o ErrorIST 2.0 irá substituir a sequência da frase pela correspondente no ficheiro. Alguns erros de ***Spelling*** já geravam alguns erros comuns, mas de forma aleatória, e utilizam o *Sounds File* para sequências como o uso de *'c'* e *'k'*, como referido na Subsecção 3.2.

Para além dos erros disponbilizados pelo ErrorIST 2.0, os utilizadores poderão completar o ficheiro com mais sequências de erros de acordo com o cenário de avaliação.

Proclise

Como apresentado anteriormente, o uso de clíticos é um erro frequente de sintaxe. Dentro dos erros com clíticos, os erros de próclise são os mais frequentes e, por esse motivo, foram introduzidos no ErrorIST 2.0 como ***Proclise Errors***.

Na próclise, o pronome pessoal é colocado antes do verbo quando é precedido de alguns advérbios, conjunções subordinativas, frases interrogativas ou exclamativas e partículas de negação.
Para introduzir este tipo de erro, são avaliadas as etiquetas morfológicas das palavras e o erro é introduzido quando temos as sequências apresentadas na Tabela 4.1:

Tabela 4.1: Próclise

Sequência	Frase Original	Frase com Erro
CONJ PRON VERB	*"O menino **que lhe chama** Joaquim."*	*"O menino ***que chama-lhe*** Joaquim."*
ADV PRON VERB	*"**Assim se resolvem** os exercícios."*	*"***Assim resolvem-se*** os exercícios."*

Wrong e Omission Errors

Para além dos erros de omissão já existentes, ***Omission Content*** e ***Omission Function***, foram introduzidos os ***Omission Determiner***, ***Omission Preposition*** e os ***Omission Pronoun***. Estes, omitem determinantes, preposições e pronomes, embora os ***Omission Function*** e os ***Omission Content*** já o fizessem de um modo mais geral. Os ***Wrong Determiner*** e os ***Wrong Preposition Errors*** identificam, através do POS *tagger*, e substituem os determinantes e as preposições da frase.

Na frase *"Obrigado por teres vindo."*, após a introdução de um ***Wrong Preposition Error***, a preposição *'por'* é substituída pela preposição *'*salvo*'*. A substituição é feita com o apoio a um ficheiro que contém todas as preposições ou, no caso dos ***Wrong Determiner Errors***, todos os determinantes.

Wrong Language Variety

Os ***Wrong Language Variety Errors*** são inseridos no ErrorIST 2.0 da mesma forma que os ***Common Errors***, ou seja, o utilizador deve fornecer um ficheiro com as sequências das palavras correspondentes à variedade da língua pretendida, para que seja feita a substituição. Este tipo de erro é um tipo de erro de variação dialetal que afeta, sobretudo, o léxico das variedades e a colocação de clíticos. Por omissão, foi aplicada a variante do Português do Brasil, mas poderá ser aplicada qualquer outra variante caso o utilizador disponibilize um ficheiro *input* com as sequências de palavras da variante pretendida. Ao introduzir este tipo de erro na frase *"Hoje fui ao **talho**."*, obtemos *"Hoje fui ao ***açougue***"*.

Register

Os ***Register Errors*** são erros de variação ao longo do discurso. Estes, recorrem ao POS *tagger* para identificar pronomes que precedem aos verbos. Tal como os erros de *Agreement Person* e *Verb Person*, é alterada a pessoa de acordo com a terminação do verbo. Na frase *"Se quiseres, eu vou contigo"*, um possível erro gerado pelo ErrorIST 2.0 seria *"Se* ****quiser****, *eu vou contigo"*.

4.1.3 Checker

O módulo *Checker* foi introduzido para solucionar problemas de geração de erros que recorriam a *Affix Files* para a criação de novas palavras, como é o caso da conjugação de verbos e concordância. Ao gerar uma nova terminação é possível que a palavra gerada não exista. Na versão ErrorIST 2.0, os erros que utilizam os *Affix Files* como recurso para a geração de palavras, podem ser verificados pelo *Checker* se o utilizador assim o entender.

Para ativar este módulo, o utilizador deverá introduzir, juntamente com os restantes ficheiros, um ficheiro de *corpora* onde o *Checker* irá recorrer para fazer a verificação da palavra.

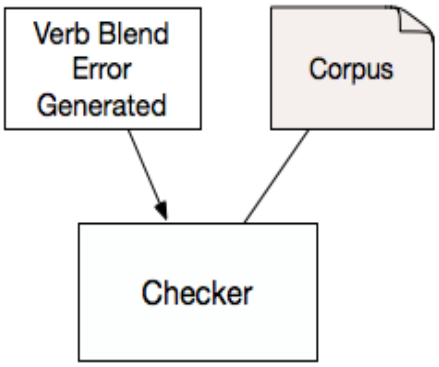

Figura 4.4: Arquitetura do módulo *Checker*

Os ***Verb Blend Errors*** serão apenas introduzidos na frase caso a palavra gerada exista no *corpus* de verificação do *Checker*.

No processo de verificação, este módulo irá receber como *input* o erro associado à verificação e um ficheiro de *corpora*, como apresentado na Figura 4.4. De seguida, verifica se a palavra gerada existe no *corpus* e, caso exista, o *Error Generator* devolve o respetivo *output*.

Consideremos a frase seguinte: *"Se eu fosse aí, trazia bolo"*. Com a versão ErrorIST 1.0 era possível gerar a palavra ***trazi***, enquanto que no ErrorIST 2.0, é gerada uma nova palavra até esta existir no *corpus* como, por exemplo, a palavra ***traziam***, referida no exemplo apresentado na Subsecção 4.1.2.

Embora a verificação seja feita, existem alguns casos cujo uso do *Checker* não evita o problema apresentado. Ao gerar o erro, o ErrorIST 2.0 pode gerar uma palavra existente no *corpus*, mas que no contexto pode não ser a mais adequada. Vejamos o seguinte exemplo onde o ErrorIST 2.0 gera um ***Verb Person Error***. Sendo *"Eu vi um urso na floresta."* a frase original e *"Eu *veste* um urso na*

floresta." a frase com o erro, verificamos que a palavra *'veste'* existe, mas não pertence ao verbo *'ver'*. No entanto, mesmo que a classe da palavra fosse verificada no *Checker*, seria identificada como verbo, que, por sua vez, coincide com a classe da palavra inicial *'vi'*.

Poderemos considerar que a utilização deste módulo tem vantagens e desvantagens. As principais vantagens da utilização do *Checker* destacam-se: no número de casos de insucesso na inserção do erro na frase; a inserção dos erros aproxima-se mais da realidade.

Por outro lado, sendo o módulo *Checker* opcional, ou seja pode ser utilizado apenas se o utilizador assim o desejar, poderia fazer sentido não ativar o *Checker* em determinados contextos como a avaliação com estrangeiros ou com crianças, em que as conjugações de verbos, como o ***"trazi"***, e concordâncias em número e género poderiam ser importantes para a avaliação.

4.1.4 Tracer

Como demonstrado na Figura 4.5, o módulo Tracer recebe como *input* o *Correction File* editado pelo avaliado, e acede ao *Error Data* onde estão guardadas as informações dos erros. O modo de identificação das alterações não foi alterado em relação à versão ErrorIST 1.0, descrita na Subsecção 4.1.4.

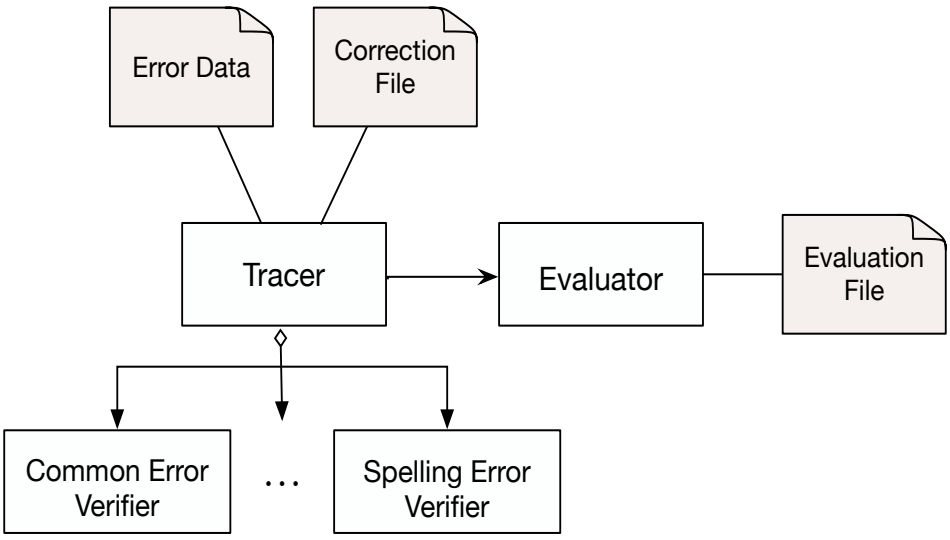

Figura 4.5: Arquitetura do Tracer e Evaluator

4.1.5 Evaluator

Feita a verificação no módulo *Tracer*, o *Evaluator* avalia o falante e devolve um ficheiro com a avaliação, indicando se a edição foi correta, incorreta ou inesperada. A avaliação manteve-se a mesma que a apresentada pelo ErrorIST 1.0, na Subsecção 4.2.1. Nesta versão, ErrorIST 2.0, o utilizador poderá fornecer um ficheiro com os pesos associados a cada tipo de erro com o seguinte formato:

```
NOME DO ERRO - PESO
CAPITALIZATION - 4
```

A atribuição de pesos aos erros será útil caso o utilizador pretenda atribuir uma classificação ao desempenho do utilizador, de modo fazer a distinção entre os vários tipos de erros e penalizar ou beneficiar o avaliado de acordo com o seu desempenho.

4.2 Recursos utilizados pelo ErrorIST 2.0

Para além dos recursos utilizados pelo ErrorIST 1.0, referidos anteriormente na Subsecçao 3.2, foi introduzida uma nova biblioteca, o SpaCy, e recorremos a ficheiros de *input* que deverão ser fornecidos pelos utilizadores.

4.2.1 POS tagger

Na nova versão do ErrorIST, versão 2.0, introduzimos o SpaCy[1], uma biblioteca de processamento de língua natural, para a atribuição de uma etiqueta morfológica às palavras. O SpaCy foi desenvolvido em Python e Cython e está disponível para 7 línguas, o que facilitou a sua integração para o português e para o inglês.

4.2.2 Ficheiros de palavras

Os ficheiros de *input* para a geração dos ***Wrong Determiners, Wrong Prepositions*** e ***Wrong Language Variety Errors*** e ***Common Errors*** devem ser disponibilizados pelo utilizador e deverão ter o seguinte formato:

```
<palavra correta> - <palavra com erro>
académica - acadêmica
```

Os ficheiros de input devem estar associados a cada um dos idiomas. Na Tabela 4.2, em anexo, são apresentados os três idiomas com os tipos de erros compatíveis e os recursos necessários à geração dos mesmos.

4.3 Como introduzir novos idiomas

O cardápio de erros do ErrorIST 2.0 está disponível maioritariamente para português, mas e se quisermos introduzir um novo idioma? Os erros podem ser divididos em dois grupos principais: erros genéricos e erros específicos.

Os erros genéricos são todos aqueles que não necessitam de um ficheiro auxiliar para a sua geração, como é o caso dos ***Keyboard, Misordering, Punctuation Omit, Spelling*** e ***Whitespace Errors***. Todos estes tipos de erros têm apenas em conta a posição do caracter ou da palavra, o que facilita a inserção deste tipo de erro em diversos idiomas que derivem do latim, grego ou cirílico.

O outro grupo de erros é um grupo mais específico, onde são necessários recursos auxiliares para gerar os erros. Podemos ainda subdividir este grupo em: erros que utilizam um ficheiro *input* com listas

[1] https://spacy.io

de palavras; erros que utilizam um POS *tagger*; erros que utilizam ambos os recursos. Na Tabela 4.2 podemos visualizar os ficheiros necessários à geração de cada um dos tipos de erro e a compatibilidade dos erros desenvolvidos no ErroIST 2.0 para os idiomas de inglês e francês.

Tabela 4.2: Cardápio de erros do ErrorIST 2.0

Tipo de erro	**PT**	**EN**	**FR**	**Ficheiro input**	**POS tagger**
Agreement Person	X	-	-	X	X
Agreement Gender	X	-	-	X	X
Agreement Number	X	-	-	X	X
Addition Content	X	-	-	X	X
Capitalization	X	X	-	-	X
Common	X	X	X	X	-
Confusion of Senses	X	-	-	X	-
Contraction	X	X	X	X	-
Graphic Accentuation	X	-	X	X	-
Keyboard	X	X	X	-	-
Misordering	X	X	X	-	-
Omission Content	X	X	-	-	X
Omission Determiner	X	X	-	-	X
Omission Pronoun	X	X	-	-	X
Omission Preposition	X	X	-	-	X
Proclise	X	-	-	-	X
Punctuation Omit	X	X	X	-	-
Punctuation Addition	X	-	-	-	X
Register	X	-	-	-	X
Spelling	X	X	X	-	-
Verb Person	X	-	-	X	X
Verb Tense	X	-	-	X	X
Whitespace	X	X	X	-	-
Word Class	X	-	-	X	X
Wrong Determiner	X	X	-	X	X
Wrong Prepositon	X	X	-	X	X
Wrong Language Variety	X	X	X	X	-

Capítulo 5

Avaliação

Neste capítulo será descrito o processo de avaliação do sistema ErrorIST 2.0 e os respetivos resultados obtidos em cada fase de avaliação. Na primeira fase da avaliação, apresentada na Secção 5.1, é testada a inserção do erro. O ErrorIST 2.0 é de seguida testado na avaliação com alunos universitários, estrangeiros e editores da empresa Unbabel, tarefas descritas nas Secções 5.2, 5.3 e 5.4, respetivamente. Por fim, na Secção 5.5, será feita uma análise e discussão dos resultados obtidos.

5.1 Avaliação da inserção do erro

A primeira fase de avaliação do ErrorIST 2.0. é relativa à inserção de erros em português, onde é avaliado se o modo de inserção do erro corresponde ao seu objetivo. Consideramos que a inserção foi bem sucedida quando, por exemplo, ao inserir um ***Verb Person Error*** no verbo *"vi"* obtemos a palavra *"*viu*"*. Caso a palavra obtida seja *"veste"*, consideramos uma inserção sem sucesso, pois não existe nenhum tempo no verbo *"ver"* que tenha a conjugação *"veste"*.

Para a avaliação, foram utilizados dois textos jornalísticos[1] para a geração dos erros. Foram gerados 8 erros de cada um dos tipos, apresentados na Tabela 5.1, sem a ativação do módulo *Checker*. De seguida, pedimos a 2 peritos, professoras universitárias, que avaliassem cada erro, analisando se este cumpria com o seu objetivo e se as palavras geradas faziam sentido. Os resultados obtidos encontram-se discriminados na Tabela 5.1.

Dos 216 erros gerados, concluímos que os erros foram inseridos com sucesso 85,18% das vezes. Devemos salientar que os restantes 14,82% são erros em que falhou a análise do *POS tagger* utilizado ou que resultaram da aplicação das regras dos ficheiros de afixos que geraram palavras inexistentes.

Observando a tabela, os ***Agreement Gender Errors*** destacam-se por um elevado número de insucessos. Ao inserir o erro na palavra *"algumas"*, através dos ficheiros de afixos, são geradas palavras que não existem ou não fazem sentido no contexto, como a palavra *"*algumos*"*, onde o género deveria ser alterado para o masculino *"alguns"*. Consideramos, assim, que *"*algumos*"* é um caso de insucesso.

[1] https://www.publico.pt/jornal e https://www.dn.pt/default.aspx, último acesso 20 de Agosto 2018

Tabela 5.1: Cardápio de erros do ErrorIST 2.0

Erro	Sucesso	Insucesso
Agreement Person	8	0
Agreement Gender	1	**7**
Agreement Number	8	0
Addition Content	7	1
Capitalization	5	**3**
Contraction	8	0
Common	8	0
Confusion of Senses	3	**5**
Graphic Accentuation	8	0
Keyboard	8	0
Misordering	8	0
Omission Content	8	0
Omission Determiner	8	0
Omission Pronoun	8	0
Omission Preposition	8	0
Proclise	7	1
Punctuation Omit	8	0
Punctuation Addition	7	1
Register	6	2
Spelling	8	0
Verb Person	6	2
Verb Tense	6	2
Whitespace	8	0
Word Class	5	**3**
Wrong Determiner	8	0
Wrong Prepositon	7	1
Wrong Language Variety	8	0
Resultado	**184**	**32**

O mesmo acontece com os ***Word Class Errors***, que utilizam o mesmo modo de geração. Como apresentado na Subsecção 4.1.2, é alterado o sufixo da palavra e, consequentemente, a sua categoria. Existem casos onde a aplicação da regra não faz sentido, por exemplo, quando a palavra *"tem"* é alterada para *"*teções*"*. Estes são casos em que a utilização de um desambiguador sintático que seja capaz de identificar sintagmas nominais e sintagmas verbais, ou até mesmo a utilização do *Checker*, poderia fazer diferença na qualidade de inserção dos erros.

Nos ***Capitalization Errors*** o número de casos de insucesso deve-se a erros na atribuição da etiqueta morfológica. Na frase *"O número da conta bancária"* a atribuição das etiquetas morfológicas é feita da seguinte forma:

`[('O', art'), ('número', 'n'), ('da', 'n'), ('conta', 'v-fin'), ('bancária', 'adj')]`

Sendo que a maiúscula é, apenas, grafada em nomes, o ErrorIST 2.0 irá selecionar como possíveis casos as palavras *"número"* e *"da"*. No entanto, a palavra *"da"* é uma contração (*de + a*) da preposição *"de"* com o artigo definido *"a"*. O ErrorIST 2.0 ao grafar a palavra *"*Da*"* com maiúscula, estará a gerar um caso de insucesso por *"*Da*"* não corresponder à classe dos nomes, na realidade. Se a palavra *"número"* fosse grafada com maiúscula já seria considerado um caso de sucesso.

Os **Confusion of Senses Errors** são gerados com recurso ao *Wiktionary* e por vezes apresentam casos de sucesso quando a palavra selecionada possuí homónimos e é traduzida de forma errada para outro idioma. Por exemplo, a palavra *"Rede"*, na frase *"Rede social."* é traduzida para inglês, *"net"*, e novamente para português e como *"net"* é uma palavra que tem outros significados associados que não correspondem à tradução, a frase obtida pelo ErrorIST 2.0 é *"*Armadilha* social."*. Assim, consideramos que o erro foi inserido com sucesso. Um caso em que não consideramos que tenha sido inserido um erro é na frase *"O nível de inglês dele é **alto**."*, em que a palavra *"alto"* é traduzida para inglês como *"loud"* e quando é novamente traduzida para português temos como traduções *"alto, elevado e intenso "*. Uma frase possível é *"O nível de inglês dele é ***intenso***."*. Na versão anterior, ErrorIST 1.0, já existia este problema na geração dos **Confusion of Senses Errors** e, nesta versão, o problema apresentado não foi resolvido, pelo que o número de casos de insucesso se manteve em relação ao ErrorIST 1.0.

A verificação da existência das palavras geradas com recurso a ficheiros de afixo era um problema existente na versão ErrorIST 1.0. Assim, para solucionar os casos de insucesso como os dos ***Agreement Gender Errors*** e os ***Word Class Errors***, ativámos o modulo *Checker* e voltámos a gerar 8 erros de cada um destes tipos e obtivemos os resultados da Tabela 5.2.

Tabela 5.2: Resultados obtidos pela ativação do módulo *Checker*

Erro	Sucesso	Insucesso
Agreement Gender	6	**2**
Word Class	8	**0**
Resultado	**14**	**2**

Com o *Checker* ativo conseguiu-se diminuir o número de casos de insucesso em ambos os erros. No entanto, os ***Agreement Gender Errors*** apresentaram dois casos de insucesso nas seguintes frases:

*"...compartimento **anexo** à moradia"*, nesta frase a palavra *"anexo"* é um substantivo e não tem feminino. Com a introdução do erro foi gerada a palavra *"*anexa*"* que existe no *corpus* de valiadação *Checker*.

Porém, *"anexa"* é uma conjugação do verbo *"anexar"*, o que torna esta inserção um caso insucesso. O mesmo ocorreu com a frase *"...os que a ela recorrem."* em que *"ela"* foi substituída por *"*elo*"*, também validada pelo *Checker*, mas que não cumpriu com o objetivo que seria substituir *"ela"* por *"ele"*.

5.2 Avaliação com alunos universitários

Avaliada a inserção do erro, na Secção anterior, seguiu-se a segunda avaliação com professoras da Faculdade de Letras da Universidade de Lisboa. Nesta fase de avaliação, os erros são avaliados de acordo com o objetivo do avaliador. De modo a selecionar e adaptar os erros a um nível universitário, foi necessário compreender que erros mais frequentes eram produzidos por alunos em contexto universitário. Na disciplina de escrita em geral não só analisam e produzem distintos géneros textuais, como auto e hetero avaliam as textualizações produzidas. Inserir erros frequentes automaticamente e avaliar a sua expressão em contexto ecológico de sala de aula é, assim, essencial, para melhor se compreender o âmbito de aplicação do ErrorIST 2.0. Tendo em conta este objetivo, procurou-se perceber que erros eram mais frequentes.

5.2.1 Estudo dos erros frequentes

O estudo foi realizado ao longo de 4 semestres e concluiu-se que os erros mais frequentes dos alunos são os erros de sintaxe, pontuação, morfossintaxe e ortografia. Na Tabela 2.1 seguinte são apresentados exemplos dos tipos de erros frequentes.

Tabela 5.3: Exemplos de erros frequentes

Tipo de Erro	Exemplo
Sintaxe	*"O carro foi **de encontro ao** autocarro."* *"O carro foi ***ao encontro do*** autocarro."*
Pontuação	*"O Bruno pode ficar no poder."* *"O Bruno *,* pode ficar no poder."*
Morfossintaxe	*"A criação dos gatos **deu** resultado."* * *"A criação dos gatos ***deram***resultado."*
Ortografia	*"quaisquer"* *"*quaisqueres*"*

Erros de sintaxe

Dentro dos erros de sintaxe ou construção frásica foi-nos pedido que déssemos mais destaque aos que envolvem clíticos. Já apresentados no sistema Correto, na Subsecção Subsecção 2.3, os clíticos são pronomes que atuam como complementos verbais e podem ser dispostos de três formas diferentes: antes do verbo (próclise); entre o verbo e as formas verbais (mesóclise); depois do verbo (ênclíse). A Tabela 5.4 apresenta as três posições do clítico e os respetivos exemplos.

No entanto, os erros de próclise são os mais frequentes e, por esse motivo, apenas esses foram introduzidos no ErrorIST 2.0.

Tabela 5.4: Posicionamento de clíticos

Clítico	Exemplo
Proclítico	*"O menino que **se** chama Joaquim."*
Mesoclítico	*"Perguntar-**lhe**-ão sobre o emprego novo."*
Enclítico	*"Ele ofereceu-**me** uma caneta."*

Erros de pontuação e morfossintaxe

Nos erros de pontuação é comum a separação entre sujeito e o predicado, onde a vírgula é colocada entre ambos. Na frase apresentada na Tabela 5.3, o sujeito *"O Bruno"* é separado do predicado *"pode ficar no poder."* por uma vírgula. Já os de morfossintaxe ocorrem frequentemente na conjugação de verbos e concordâncias em número, género e pessoa.

Erros de ortografia

Na categoria dos erros ortográficos, para além da acentuação gráfica e da troca de letras, analisámos as mudanças causadas pela introdução do novo Acordo Ortográfico (AO). A introdução do novo Acordo Ortográfico é, atualmente, um contributo para a maioria dos erros ortográficos na língua portuguesa, pois veio aproximar a língua escrita da língua falada, priveligiando a fonética. As principais alterações com a introdução do novo Acordo Ortográfico foram:

- Supressão das consoantes *'c'* e *'p'* quando estas não são articuladas;
  Antes do AO: *excep**p**to* / depois do AO: *exceto*
- Revisão das regras de grafar maiúsculas;
  Antes do AO: ***O**utubro* / depois do AO: *outubro*
- Revisão das regras de utilização do hífen;
  Antes do AO: *há **-** de* / depois do AO: *há de*
- Supressão de acentos maioritariamente em palavras graves;
  Antes do AO: *pára* / depois do AO:*para*

No artigo da revista Politecnia (Alexandre, 2012) são apresentadas as diferenças entre o antigo e o novo Acordo Ortográfico.

Erros de variação dialetal e de registo

A constante mudança da língua é uma consequência da transformação da sociedade ao longo do tempo e tem a contribuição de diversos fatores. Dentro do mesmo país podemos ouvir falar a língua de diferentes formas. A variação ocorre substancialmente no vocabulário, na fonética e na sintaxe. Todas

estas mudanças poderão depender da época, da região geográfica, da faixa etária, do estatuto social e do ambiente envolvente. Este tipo de variações (Mota, 2002) pode ser dividido em:

- **Variações Diatópicas** – representadas pela diversidade linguística regional ou geográfica.

  Exemplo: *'imperial'* (Lisboa) e *'fino'* (Porto).

- **Variações Diafásicas** – relacionadas com o contexto comunicativo, ou seja, com a formalidade de um texto de acordo com a situação;

  Exemplo: e-mail para um amigo (informal) e um e-mail de candidatura a uma empresa (formal).

- **Variações Diastráticas** – referem-se às diferenças entre a linguagem utilizada em grupos sociais;

  Exemplo: linguagem utilizada pelos médicos e linguagem utilizada pelos grupos de jovens.

Os ***Wrong Language Variety Errors*** são erros de variações diatópicas e são frequentes em textos traduzidos para português onde existem variações entre português europeu (PE) e português do Brasil (PB). Como exemplo, temos a palavra *"casaco"* que pode ser encontrada como: *blazer, blêizer, casaco, casaquinho, casaquinha, manteau, mantô, paletó, paletot* (da Silva, 2010). A Tabela 5.5 apresenta outros exemplos de palavras com o mesmo significado.

Tabela 5.5: Variações do vocabulário

Português Europeu	**Português Brasil**
registo	registro
aceder	acessar
equipa	time
apagar	deletar
descarregar	baixar
aplicação	aplicativo

Na construção sintática encontramos, também, alguns aspetos que se destacam como diferenciadores do PE e PB, como apresentado na Tabela 5.6.

- O pronome é colocado depois do verbo em PE e colocado antes do verbo em PB;
- O uso da contração *'à'* (*'a' + 'a'*) em PE e o uso da contração *'na' 'em' + 'a'* em PB;
- Aplicação do gerúndio como forma nominal verbal em PB, invés do verbo no infinitivo precedido de uma preposição em PE.

Os erros causados por variações diafásicas, são apresentados pela taxonomia MQM, na Subsecção 2.1.2, como ***Register Errors***.

Tabela 5.6: Variações do Português Europeu e Português do Brasil

Variação	Português Europeu	Português do Brasil
Colocação do pronome	*"Ele ofereceu-**me** um gato."*	*"Ele **me** ofereceu um gato."*
Utilização da preposição em/a	*"Vou **à** casa da avó."*	*"Vou **na** casa da avó."*
Uso do gerúndio	*"Estou **a preparar** o lanche."*	*"Estou **preparando** o lanche."*

5.2.2 Avaliação

Analisados os tipos de erros mais frequentes dos alunos, seguimos para a avaliação dos mesmos com alunos universitários.

A avaliação foi feita com 14 alunos e os erros que se adequaram melhor ao tipo de frases a avaliar, de acordo com as professoras, foram os ***Commons***, ***Graphic Accentuation***, ***Proclise Errors*** e ***Punctuation Omit***. Por cada tipo de erro foram gerados 10 erros e, estes, foram inseridos em textos jornalísticos.

Das inserções bem-sucedidas, apresentadas na Tabela 5.1, as professoras selecionaram os erros que se adequavam aos alunos. Em seguida, foram dados aos alunos os textos com 17 erros, selecionados pelas professoras, para que estes os corrigissem, não tendo qualquer informação sobre o tipo de erro nem quantos erros existiam por frase. Depois das intervenções dos alunos, o ErrorIST 2.0 fez a avaliação de acordo com as métricas aplicadas pelo *Evaluator*, descritas na Subsecção 4.2.1. Na Tabela 5.7 seguinte, estão representados os resultados atribuídos pelo ErrorIST 2.0 às edições.

Tabela 5.7: Avaliação com alunos obtida pelo ErrorIST 2.0

Erro	Nº Erros Inseridos	Edição Correta	Edição Incorreta	Edição Inesperada	% Erros Dispensados
Common	56	18	6	32	42,85%
Graphic Accentuation	42	11	7	24	75%
Proclise	28	13	3	12	57,14%
Punctuation Omit	112	38	12	62	44,64%
Total	**238**	**80**	**28**	**130**	**45,38%**

A Tabela 5.7 apresenta, também, a percentagem de erros que foram dispensados de intervenção manual na atribuição da avaliação em cada tipo de erro (última coluna), dada pela expressão:

*((Edições Corretas + Edições Incorretas)/Nº de Erros Inseridos) * 100*

Nos ***Common Errors*** e nos ***Punctuation Errors***, mais de metade das edições foram inesperadas (`INDEF`), ou seja são correções que o programa não conseguiu avaliar como corretas (`OK`) ou incorretas `KO` e que necessitam de verificação humana.

Relembramos que os casos que exigem intervenção humana são casos como, por exemplo, o aluno não modificar o erro e fazer outras alterações à frase. O número elevado de casos inesperados po-

derá ser influenciado pelo tipo de erro. Uma das frases para avaliação de um ***Common Error*** era *"Num dos casos, em 2007, Gimenez fez* ***querer*** *à pessoa que..."*. Apenas 3 alunos detetaram o erro, substituindo-o corretamente por *'crer'*. Os outros alunos editaram a frase de outras formas sem detetar o erro, o que fez com que 11 casos fossem inesperados.

Uma das motivações do ErrorIST 2.0 é a diminuição do número de intervenções humanas na verificação das correções e, desta forma, foram evitadas 45,38% de intervenções.

As classificações que careceram de validação manual foram analisadas por uma das professoras que classificou as edições como corretas ou incorretas e obtiveram-se os seguintes resultados da Tabela 5.8:

Tabela 5.8: Verificação das edições inesperadas

Erro	**Edição Inesperada**	**Edição Correta**	**Edição Incorreta**
Common	32	6	26
Graphic Accentuation	24	3	21
Proclise	12	2	10
Punctuation Omit	62	33	29
Total	**130**	**44**	**86**

Os erros adaptados a esta avaliação são erros complexos e que exigem verificação humana, em muitos dos casos, por se tratar de uma avaliação mais criteriosa. Para os erros de pontuação, foi necessária verificação humana, pois, ao inserir ou omitir uma vírgula para além da pretendida, a frase poderá manter-se correta e o ErrorIST 2.0 não conseguir avaliar a intervenção como tal.

5.3 Avaliação com estrangeiros

Selecionámos um dos novos idiomas, o francês, para avaliar a ferramenta com 3 falantes fluentes em francês com idade e nível de escolaridade diferente. Começámos por preparar a avaliação, recorrendo a erros comuns da língua existentes no *Language Tool*, apresentado na Secção 2.3, e outros documentos de estudo dos erros em francês (Strube Den Lima, 1990). Foram gerados 10 erros de cada um dos tipos ***Common***, ***Contraction***, ***Graphic Accentuation*** e ***Spelling***, selecionados pelo perito com base no estudo citado.

De seguida, um perito fluente em francês, externo à edição dos textos, avaliou cada erro gerado. Sendo que os erros selecionados não dependem de ficheiros de afixo e POS *tagger*, o modo de inserção foi um sucesso em todos os casos. Destes erros inseridos, o perito selecionou 14 erros para avaliação.

O texto foi dado aos falantes para que o corrigissem, não tendo informações sobre os tipos de erros; apenas foram informados de que existia um erro por frase.

Tendo em conta que a avaliação foi mais simplista e com apenas 3 participantes, o número de

casos inesperados foram menores. Mais uma vez, após a edição dos falantes, o ErrorIST 2.0 fez a classificação das edições, tal como apresentado na Tabela 5.9.

Tabela 5.9: Avaliação dos falantes fluentes em francês atribuída pelo ErrorIST 2.0

Erro	Nº Erros Inseridos	Edição Correta	Edição Incorreta	Edição Inesperada	% Erros Dispensados
Common	18	14	1	3	83,33%
Contraction	6	2	1	3	50%
Graphic Accentuation	12	9	1	2	83,33%
Spelling	6	4	0	2	66,67%
Total	**42**	**29**	**3**	**10**	**76,19%**

Analisando os resultados correspondentes à avaliação do ErrorIST 2.0, conclui-se que o ErrorIST 2.0 conseguiu avaliar 76,19% dos erros como corretos ou incorretos.

Os restantes 23,81% correspondem aos 10 casos inesperados que foram seguidamente avaliados pelo perito para que fossem classificados, igualmente, como corretos ou incorretos.

Um dos casos inesperados foi um ***Contraction Error***, onde os avaliados não detetaram o erro, e alteraram o contexto da frase sem editar o erro como esperado. Na frase original *"Je vais **à la** pharmacie."(Eu vou à farmácia.)* foi inserido um ***Contraction Error*** gerando a frase *"Je vais ***au*** pharmacie."*, e os falantes corrigiram para *"Je vais au **gym**."(Eu vou ao ginásio.)*. Sendo que este tipo de erro poderá ter mais do que uma opção de correção, é necessário ser verificado por um humano.

5.4 Avaliação com editores da Unbabel

A avaliação de editores, como foi referido anteriormente na Secção 1.1, é uma tarefa dispendiosa para os avaliadores. Como proposta de solução ao problema, apresentado pela Unbabel, analisámos os tipos de erros mais frequentes dos editores em conjunto com um perito da Unbabel e introduzimos novos tipos de erros ao ErrorIST 2.0 – ***Wrong Preposition, Wrong Determiner, Omission Preposition, Omission Determiner, Wrong Language Variety, Register, Whitespace*** e ***Graphic Accentuation*** – descritos na Subsecção 4.1.2.2.

Para a avaliação com editores foram gerados 10 erros de cada um dos tipos, em português, disponíveis no cardápio do ErrorIST 2.0, e inseridos em textos com formato de e-mail. Foi feita uma pré-validação dos tipos de erros gerados e selecionados os que mais se adequavam aos textos, dos quais foram selecionados os erros da Figura 5.1.

Até ao momento, ainda não nos foi possível obter os resultados dos erros gerados, pois estão a ser validados por peritos da Unbabel. Após esta validação, o objetivo do nosso trabalho é concluir a avaliação com a interação dos editores da Unbabel e fazer um balanço da utilidade do ErrorIST 2.0 neste tipo de contexto.

5.5 Discussão

Terminada a avaliação e observando as tabelas anteriores, podemos concluir que o ErrorIST 2.0 foi capaz de avaliar como correto ou incorreto 45,38% dos erros na avaliação com alunos e 76,19% na avaliação com falantes fluentes em francês.

Na avaliação com alunos de Produção de Português Escrito, os ***Common Errors*** e os ***Punctuation Errors*** destacaram-se por terem um número de casos inesperados superior a 50%. No caso dos ***Punctuation Errors*** o número de casos inesperados, em percentagem, é maior do que a percentagem apresentada pelos ***Common Errors***. Sendo os ***Punctuation Errors*** introduzidos por omissão, o ErrorIST 2.0, e não contendo um desambiguador sintático, não consegue identificar sintagmas nominais e verbais para desambiguar casos em que as vírgulas foram adicionadas ou omitidas corretamente e acaba por retornar indefinido. A percentagem de casos inesperados na avaliação com estrangeiros foi menor, apresentando apenas 23,81%. A prestação do ErrorIST 2.0 permitiu que dos 42 erros, 10 tivessem de ser verificados manualmente.

Na Figura 5.1 podemos visualizar o cardápio de erros disponibilizados pelo ErrorIST 2.0 e os erros selecionados para cada umas das avaliações realizadas pelo ErrorIST 2.0. O cardápio de erros do ErrorIST 2.0 permite adaptar o sistema a diferentes cenários de avaliação tendo em conta as necessidades apresentadas pelos peritos, como demonstrado nas avaliações realizadas.

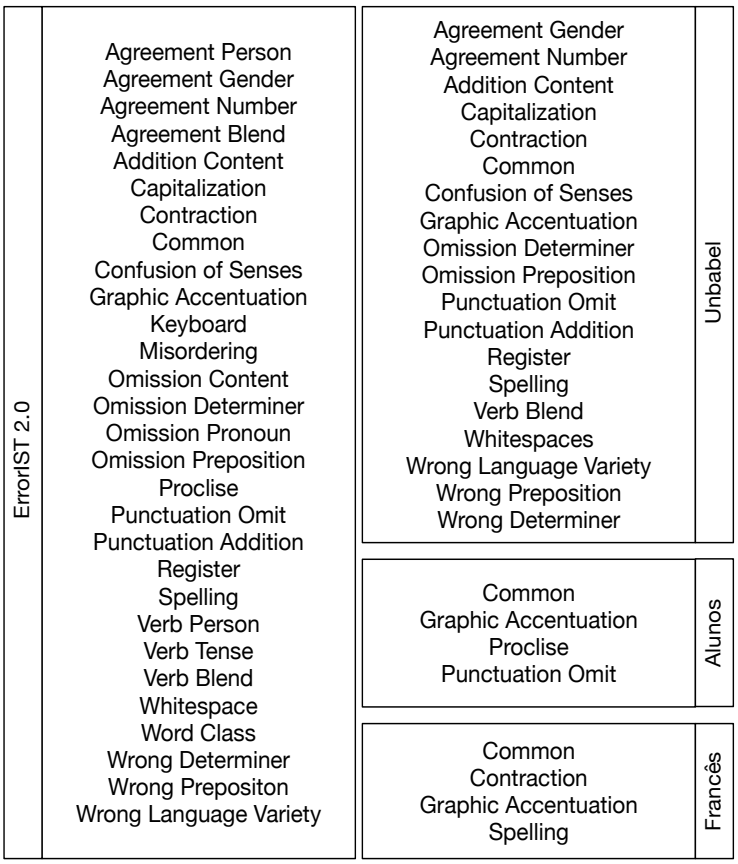

Figura 5.1: Cardápio ErrorIST 2.0 e erros selecionados para cada contexto

Capítulo 6

Conclusões e Trabalho Futuro

6.1 Conclusões

O ErrorIST 2.0 apresenta melhorias em relação à versão ErrorIST 1.0, anteriormente apresentada. O ErrorIST 2.0 tinha como objetivos principais: a melhoria do modo de geração dos erros de grafar a maiúscula ou minúscula, pontuação e conjugaçãos de vernos; aumentar o número de erros disponibilizados; introduzir erros de discurso; introduzir novos idiomas; avaliar o sistema em novos cenários.

Nos ***Capitalization Errors***, o grafar de maiúscula passou a ser feito apenas em palavras cujo o POS *tagger* identifica como nomes. Os ***Punctuation Errors*** faziam a inserção de qualquer tipo de caracter e foram limitados aos caracteres principais de pontuação. Para solucionar a geração de palavras inexistentes foi adicionado um módulo de verificação à arquitetura do sistema, o módulo *Checker*, que poderá ser ativo se o utilizador assim o desejar. A introdução do módulo *Checker* diminuiu o número de insucessos na inserção dos erros que utilizam os ficheiros de afixo na sua geração, como os ***Verb Errors***, ***Agreement Errors*** e os ***Word Class Errors***.

Em comparação aos sistemas de geração GenERRate e Missplel, e ainda à versão anterior, o ErrorIST 2.0 disponibiliza um cardápio de erros mais completo e adaptável a diferentes tipos de avaliação. O cardápio disponibiliza erros genéricos, que não necessitam de recursos auxiliares, e, também, erros mais específicos de um dado idioma. Para além de erros em português, o ErrorIST 2.0 disponibiliza alguns tipos de erros para inglês e francês. Para a introdução de novos tipos de erros no ErrorIST 2.0, foram tidos em conta dados estatísticos disponibilizados por peritos em cada um dos cenários – avaliação com alunos fluentes em português, fluentes em francês e editores da Unbabel – de avaliação.

O objetivo principal do ErrorIST 2.0 é minimizar a tarefa de validação humana e, apesar do elevado número de casos indefinidos, o ErrorIST 2.0 conseguiu diminuir a tarefa manual em mais de 45% dos erros. Na avaliação com alunos fluentes em português, o ErrorIST 2.0 dispensou de verificação manual 45,38% dos erros e na avaliação com fluentes em francês conseguiu avaliar automaticamente 76,19% dos erros. Estes resultados têm vários fatores envolvidos como os tipos de erros gerados, o nível de conhecimento dos avaliados e ainda a verificação feita pelo ErrorIST 2.0, que poderá ser melhorada futuramente. Com as melhorias realizadas na versão ErrorIST 2.0 e com o cardápio de erros

disponibilizado pelo ErrorIST 2.0, na Figura 5.1, concluímos que é possível automatizar grande parte da avaliação dependendo do tipo de erros e do cenário de avaliação, apesar de haver uma percentagem que necessita sempre de verificação humana.

6.2 Trabalho Futuro

O trabalho desenvolvido cumpre com os objetivos inicialmente apresentados, contudo existem pontos que podem ser melhorados no futuro que contribuirão para um melhor desempenho e usabilidade do ErrorIST 2.0.

Dentro dos problemas encontrados ao longo do desenvolvimento da ferramenta, começamos por referir os problemas na geração de erros e as suas propostas de resolução.

Os ***Confusion of Senses Errors*** foram mencionados na versão anterior como um aspeto a melhorar, porém, continuam a ser um tipo de erro difícil de inserir devido à utilização do *Wiktionary* e por falta de avaliação do contexto da frase, como por exemplo na frase *"Obrigada pelo seu **tempo**. "tempo"* ser substituído por *"ponto"*.

A inserção de erros em verbos falha em muitos casos, pois a distinção entre verbos regulares e irregulares não é linear e, ao aplicar as regras para a alteração do sufixo, acabam por ser geradas terminações inexistentes, aumentando o número de erros a serem verificados pelos peritos. A geração de um *Verb Tense* na frase *"Ainda vai **ser** vencedor."* tem como resultado possível a alteração do verbo *"ser"* para *"si"*.

O *corpus* de verificação do módulo *Checker* poderá ser alargado através da inserção de novas palavras.

Os ***Wrong Language Variety*** e os ***Register Errors*** poderiam tirar partido de ficheiros adicionais, como os ficheiros de afixo, ou da análise com n-gramas.

Para finalizar, foram adicionados ao ErrorIST 2.0 erros noutras línguas: inglês e francês. Os erros de inglês recorrem a um POS *tagger* para a atribuição de etiquetas morfológicas específicas do inglês, mas é necessário inserir mais regras para que mais erros mais específicos da língua sejam suportados, consiguindo-se assim uma avaliação mais fina.

Os erros em francês foram inseridos e gerados apenas com recurso a ficheiros auxiliares, mas é necessário a integração de um POS *tagger* para permitir que mais tipos de erros fossem inseridos, para além dos ***Common, Wrong Preposition, Graphic Accentuation*** e outros erros mais genéricos como os de ***Spelling, Misordering, Punctuation Omit*** e ***Keyboard***.

No processo de verificação da correção, de modo a dispensar mais casos indefinidos, o ErrorIST 2.0 deveria recorrer a dicionários com sinónimos para resolver casos em que o editor altera a palavra para o seu sinónimo. A ambiguidade é, também, um dos problemas no processo de verificação das intervenções dos avaliados. Alguns dos casos indefinidos poderiam ser resolvidos com recurso a um desambiguador sintático como os casos em que o avaliado editou a frase corretamente, mas não da forma esperada. Como exemplos temos a colocação das vírgulas nos ***Punctuation Errors*** e a utilização de palavras sinónimas na correção de um erro.

Anexo A

Taxonomias de Erros

A.1 Erros implementados pelo ErrorIST 1.0 com base na taxonomia do L2F

Tabela A.1: *Orthography errors*

Tipo de Erro	Descrição	Exemplo
Punctuation	Uso incorreto de pontuação.	Quero ovos, farinha e açúcar. Quero ovos ***.*** farinha e açúcar.
Capitalization	Grafar de maiúscula/minúscula indevidamente	O João está pronto. O ***joão*** está pronto.
Spelling	Troca, remoção ou adição de um ou mais caracteres na palavra.	A Susana veio buscar os bolos. A Susana veio buscar os ***bols***.

Tabela A.2: *Discourse errors*

Tipo de Erro	Descrição	Exemplo
Style	Repetição de palavras redundantes.	*... autorização para permitir...* *...****permissão**** *para permitir...*
Variety	Inconsistência no texto devido a variações do idioma. (ex.:Inglês para Português(Brasil))	*In your room. (No seu quarto.)* ****Em**** *seu quarto.*
Should not be translated	Tradução de expressões ou nomes que não devem ser traduzidos.	*Gostas do Mr. Bean?* *Gostas do* ****Senhor Feijão****?*

Tabela A.3: *Grammar and Semantic errors*

Tipo de Erro		**Descrição**	**Exemplo**
Misselection	Verb Tense	Tempo verbal incorreto.	*Ele tinha comprado uma mala.* *Ele ***ter*** comprado uma mala.*
	Verb Person	Conjugação do verbo de acordo com a pessoa incorreta.	*A Maria é fã da Isabel.* *A Maria ***são*** fã da Isabel.*
	Verb Blend	Tempo verbal e/ou da pessoa incorreto.	*Ele comeu um bolo.* *Ele ***comeriam*** um bolo.*
	Agreement Gender	Concordância em género incorreta.	*A Sara correu atrás da vaca.* *A Sara correu atrás ***do*** vaca.*
	Agreement Number	Concordância em número incorreta.	*... sempre uma lição de moral.* *...sempre uma ***lições*** de moral.*
	Agreement Person	Concordância em pessoa incorreta.	*As crianças ***pode*** cair.*
	Agreement Blend	Concordância em género, número ou pessoa incorreta.	*Viste aqueles golos?* *Viste ***aquela*** golos.*
	Contraction	O uso da contração é substituído por uma preposição e um determinante.	*Num berço.* ****Em um*** berço.*
Misordering		Trocas entre palavras na frase.	*Quero comer pão.* *Quero ***pão*** comer.*
Confusion of Senses		Tradução de uma palavra que não se enquadra no contexto da frase.	*I see well with glasses.* *(Eu vejo bem com os óculos.)* *Eu vejo bem com os ***copos*.***
Wrong Choice		A tradução da palavra original não corresponde à palavra traduzida.	*in the same **quarter**.*(no mesmo bairro) *no mesmo ***histórica****
Collocational		Conjunto de palavras traduzidas individualmente e não como um todo.	*“high wind”(vento forte)* *vento alto*
Idiom		Expressões idiomáticas traduzidas à letra.	*Itś raining cats and dogs.* *(“Está a chover a potes.”)* *Estão a chover cães e gatos.*
Variety		Estruturas lexicais ou gramaticais associadas a uma variação da língua. (ex.: brasileiro)	*In his room. (No seu quarto)* ****Em seu**** *quarto.*
Should not be translated		Palavras no texto original, como nomes, que são traduzidas.	*He is Peter.* *Ele é o ***Pedro****.

Tabela A.4: *Lexical errors*

Tipo de Erro		**Descrição**	**Exemplo**
Content Word	Omission	Remoção de palavras que carregam informação na frase (ex.:nomes,adjetivos)	No sábado foram ver as luzes de Natal. No sábado ** ver as luzes de Natal.
	Addition	Adição de palavras que carregam informação na frase (ex.:nomes, adjetivos)	... Amazónia é fantástica. ...Amazónia é fantástica ***linda***.
Function Word	Omission	Remoção de palavras que expressam relações entre as palavras (ex.: preposições, pronomes)	Essa era a rapariga que fugiu. Essa era a rapariga ** fugiu.
	Addition	Adição de palavras que expressam relações entre as palavras (ex.: preposições, pronomes)	Eu estou sempre a viajar. Eu estou ***a*** sempre a viajar.
Untranslated		Conteúdo no texto traduzido que não foi traduzido.	...world of fashion. ...mundo da ***fashion***(moda).

A.2 Erros da taxonomia MQM

Tabela A.5: Dimensão *Accuracy*

Tipo de Erro		**Descrição**
Addition		Adicionar uma palavra ao texto.
Omission		Remover uma palavra do texto.
Mistranslation	Ambiguous Translation	Traduções ambíguas de textos traduzidos.
	Date/Time	Datas ou horários que não correspondem ao texto original.
	Entity	Entidades que não correspondem ao texto original.
	False Friend	Tradução de uma palavra noutra semelhante, mas errada no texto traduzido.
	Technical Relationship	Tradução incorreta de relacionamentos em conceitos técnicos.
	Number	Inconsistências de número entre o texto original e a sua tradução.
	Overly Literal	Traduções literais.
	Should have not been translated	Conteúdo que não deve ser traduzido.
	Unit Conversion	Medidas de conversão incorretas entre o texto original e a sua tradução.

Tabela A.6: Dimensão *Style*

Tipo de Erro	**Descrição**
Register	O texto não está de acordo com a formalidade.
Register (variants/slang)	Linguagem inapropriada
Awkward	O texto é afetado por uma grande quantidade de fenómenos que prejudicam a leitura e a sua compreensão.(ex.: uso de cláusulas)
Company style	Não estão de acordo com as diretrizes da empresa.
Inconsistent style	Inconsistência entre as secções do texto (ex.: uma secção concisa e outra abstrata)
Third-party style	As especificações estipuladas não são seguidas. (ex.: seguir um guia linguístico)
Unidiomatic	Traduções gramaticais, mas não idiomáticas.

Tabela A.7: Dimensão *Terminology*

Tipo de Erro	**Descrição**
Inconsistent with termbase	O termo é utilizado incorretamente de acordo com a terminologia apresentada na base de dados.
Inconsistent with termbase (Company terminology)	Viola as diretrizes de terminologia da empresa.
Inconsistent with termbase (Third-party)	Viola as diretrizes de terminologia de terceiros.
Inconsistent with domain	Um termo é usado contra as especificações do domínio.
Inconsistent use of terminology	A terminologia utilizada de forma inconsistente com o texto.
Inconsistent use of terminology (Multiple terms for concept in source)	O mesmo termo é referido de formas diferentes.
Inconsistent use of terminology (Multiple translations for the same term)	Vários termos para o mesmo conceito no texto original e os respetivos conceitos são traduzidos.

Tabela A.8: Dimensão *Fluency*

Tipo de Erro	**Descrição**
Spelling	Troca, remoção ou adição de um ou mais caracteres na palavra.
Capitalization	Palavra iniciada com letra maiúscula indevidamente e vice-versa
Diacritic	Acentuação incorreta de um caracter.
Cohesion	Ligação incorreta entre as várias partes de um texto.
Inconsistency	Uma entidade é referenciada com diferentes expressões num texto.
Grammatical-Register	Problemas de formalidade na linguagem.
Typography	Problemas na apresentação de um texto. (ex.: *Punctuation Error*)

Referências

N. Alexandre. O acordo ortográfico: o que muda. *Politecnia - Instituto Politécnico de Lisboa*, março 2012.

J. Bigert, L. Ericson, and A. Solis. Missplel and autoeval: Two generic tools for automatic evaluation. In *Proceedings of Nodalida*. Nordic Conference of Computational Linguistics, 2003.

S. Bird. Nltk: the natural language toolkit. In *Proceedings of the COLING/ACL on Interactive presentation sessions*. Association for Computational Linguistics, 2006.

O. Bojar. Analyzing error types in english-czech machine translation. *The Prague Bulletin of Mathematical Linguistics*, 2011.

Â. Costa, W. Ling, T. Luís, R. Correia, and L. Coheur. A linguistically motivated taxonomy for machine translation error analysis. *Machine Translation*, 2015.

A. da Silva. Measuring and parameterizing lexical convergence and divergence between european and brazilian portuguese. *Advances in cognitive sociolinguistics*, 2010.

F. J. Damerau. A technique for computer detection and correction of spelling errors. *Commun. ACM*, 1964.

T. M. F. dos Santos. Errorist: towards the automatic evaluation of editors. *Master's thesis, Instituto Superior Técnico-Universidade Técnica de Lisboa*, 2016.

M. Felice and Z. Yuan. Generating artificial errors for grammatical error correction. In *EACL*, 2014.

O. E. Foster, Jennifer e Andersen. Generrate: Generating errors for use in grammatical error detection. In *Proceedings of the Fourth Workshop on Innovative Use of NLP for Building Educational Applications*. Association for Computational Linguistics, 2009.

Liţă, L and Ittycheriah, A and Roukos, S and Kambhatla, N. tRuEcasIng. In *Proceedings of ACL*, 2003.

A. Lommel, H. Uszkoreit, and A. Burchardt. Multidimensional quality metrics (mqm): A framework for declaring and describing translation quality metrics. *Tradumàtica*, 2014.

J. C. Medeiros. Processamento morfológico e correcção ortográfica do português. *Master's thesis, Instituto Superior Técnico-Universidade Técnica de Lisboa, February*, 1995.

A. Mendes, S. Antunes, M. Janssen, and A. Gonçalves. The cople2 corpus: a learner corpus for portuguese. In *LREC*, 2016.

J. Mota. A variação diafásica no português do brasil. *Revista de Letras*, 2002.

O. G. Pereira. Erro humano: uma conferência internacional. *Análise psicológica*, 1983.

J. Rasmussen. Human errors. a taxonomy for describing human malfunction in industrial installations. *Journal of occupational accidents*, 1982.

V. L. Strube Den Lima. *A contribution to the study of error treatment at lexical-syntactical level in a french written text*. PhD thesis, Mar 1990.

D. Vilar, J. Xu, L. D'haro, and H. Ney. Error Analysis of Statistical Machine Translation Output. European Language Resources Association (ELRA), May 2006.

J. L. Zorzi. As trocas surdas sonoras no contexto das alterações ortográficas. *Revista SOLETRAS*, (15), 2008.