

Linguagem Controlada entre Português e Chinês

尹雪璐 (Yin Xuelu)
乔建珍 (Qiao Jianzhen)

Resumo: No século XX, com a globalização da economia, a crescente migração e o desenvolvimento da tradução automática, o conceito de linguagem controlada surgiu e continuou a expandir-se e aprofundar-se, atraindo ampla atenção de estudiosos da linguística e dos círculos de tradução. Com esse texto, objetivamos fazer uma reflexão sobre a linguagem controlada, destacando sua origem, tipologias e função. Além disso, evidentemente, refletimos brevemente sobre a Linguagem Controlada (LC) no que diz respeito à língua chinesa e à língua portuguesa. Por fim destacamos a relevância de tal estudo, da investigação sobre essa linguagem, considerando a comunicação cada vez mais estreita entre a China e os países de língua portuguesa. Tal linguagem poderá desempenhar um papel importante no domínio da tradução automática, do ensino, da redação de textos para comunicação técnica e no domínio da inteligência artificial, evitando ambiguidades, favorecendo a compreensão, a traduzibilidade.

Palavras-chave: Linguagem controlada. Língua chinesa e língua portuguesa. Tradução.

Controlled Language between Portuguese and Chinese

Abstract: In the 20th century, with the globalization of the economy, increasing migration and the development of automatic translation, the concept of controlled language emerged and continued to expand and deepen, attracting wide attention from scholars in linguistics and translation circles. With this text, we aim to reflect on controlled language, highlighting its origin, typologies and function. In addition, of course, we briefly reflect on Controlled Language (CL) with regard to the Chinese language and the Portuguese language. Finally, we highlight the relevance of such a study, of the research on this language, considering the increasingly close communication between China and Portuguese-speaking countries. Such language could play an important role in the field of automatic translation, teaching, writing texts for technical communication and in the field of artificial intelligence, avoiding ambiguities, favoring understanding and translatability.

Keywords: Controlled language. Chinese language and Portuguese language. Translation.

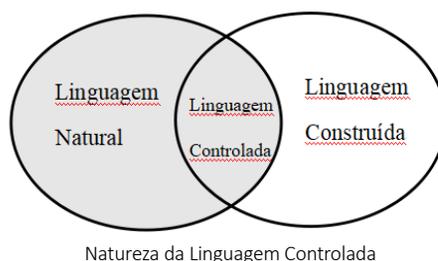
No século XX, com a globalização da economia, a crescente migração e o desenvolvimento da TA, o conceito de linguagem controlada surgiu e continuou a expandir-se e aprofundar-se, atraindo ampla atenção de estudiosos da linguística e dos círculos de tradução.

1. A Linguagem Controlada: conceito e aplicação

O conceito de LC, também conhecida por linguagem natural controlada, foi originalmente proposto, em 1930, por linguistas e estudiosos que procuraram estabelecer uma variedade “mínima” de inglês, com o fim de o tornar acessível e utilizável internacionalmente por um maior número de pessoas (especialmente falantes não-nativos) (Schwiter, 2007). A LC é uma forma especial de linguagem que se situa entre a linguagem natural e a linguagem construída. Os humanos usam a linguagem natural, tratando-se da forma mais comum de comunicação, à qual estamos acostumados desde que nascemos. A LC é um tipo especial de linguagem natural, pensado e expresso de uma forma concisa, precisa e clara, sendo também uma linguagem planeada, propositalmente pensada e processada. “*Natural language is any language developed naturally by humans, in an unpremeditated way, as a result of the innate facility for language possessed by the human intellect*” (Chomsky, 1963, p. 269).

Constructed languages (or artificial languages or planned languages) are languages that did not emerge naturally but have been consciously defined. In this broad sense, the term includes (but is not limited to) languages such as Esperanto, programming languages, and controlled natural languages (Tobias Kuhn, 2014, p. 125).

Figura 1



Pode-se notar que a linguagem natural é o estado mais natural da linguagem não processada, a linguagem construída é uma linguagem projetada com certo objetivo e a LC é uma forma de linguagem entre as duas, possuindo atributos naturais e traços de processamento.

A LC é a linguagem natural processada que tem propriedades tanto naturais como não-naturais. A naturalidade refere-se à linguagem baseada na linguagem natural, enquanto a não naturalidade significa que é produzida por meio de restrição e processamento. A naturalidade é uma propriedade essencial da LC, o que significa que esta precisa de estar em conformidade com as convenções da linguagem natural. “A linguagem natural controlada é um subconjunto da linguagem natural” (Fuchs & Schwitter 1995, p. 1). A desnaturalidade é uma propriedade geral da LC, principalmente porque a LC não é exatamente a mesma que a linguagem natural, por ter sido submetida a processamento. Os usuários de linguagem natural não consideram as expressões após o controlo efetuado na LC como sendo as mais utilizadas e mais expressivas, mas sim as mais precisas, inequívocas, concisas e fáceis de entender.

A controlled natural language is a constructed language that is based on a certain natural language, being more restrictive concerning lexicon, syntax, and/or semantics, while preserving most of its natural properties (Kuhn, 2014, p. 123).

O valor da LC reside na sua capacidade de controlar uma linguagem natural infinitamente variável e torná-la numa linguagem regular com uma estrutura clara e expressão precisa e contida, geralmente usada na redação dos documentos técnicos.

A linguagem controlada ou linguagem natural controlada é uma certa restrição ao infinito de linguagem natural, para melhorar a legibilidade, fazer o texto facilmente traduzível, eliminar a ambiguidade e reduzir a complexidade na linguagem dos textos (Schwitter, 2007, p. 3).

Como forma especial de género, os textos de comunicação técnica são materiais para a comunicação de tecnologias, produtos, segurança e processos, entre outros. Tal comunicação inclui intercâmbios nacionais e

internacionais e, para melhor realizar o objetivo da comunicação para diferentes públicos, existem certos requisitos de exatidão e concisão das expressões linguísticas que constituem uma importante base e orientação para a preparação de textos de comunicação técnica, devido à elevada legibilidade, à facilidade de compreensão e tradução, à eliminação da ambiguidade e outros aspetos da LC.

2. Desenvolvimento da Linguagem Controlada

O desenvolvimento da LC passou por três fases ou tendências: LC para entender melhor o inglês, em contexto de aprendizagem desta língua; LC para facilitar a elaboração dos manuais técnicos; e LC para beneficiar a TA.

A primeira geração foi entre as décadas de 1930 e 1980, período em que o inglês já se tornava numa língua global. A LC foi usada para entender melhor o inglês e para facilitar o ensino dessa língua aos alunos com línguas maternas diversas. O linguista britânico Charles Kay Ogden propôs e projetou o Inglês Básico, na década de 1930, no livro *Basic English: A General Introduction with Rules and Grammar*, incluindo um conjunto de uma dezena de regras de gramática inglesa e um glossário de 850 palavras. Mais tarde, a Lei dos Direitos Cívicos de 1964 estipula no seu Artigo 6.º que “todas as pessoas nos Estados Unidos, com proficiência limitada em inglês, têm o direito de receber assistência linguística nos serviços governamentais”. Esta lei reforçava a necessidade de se disponibilizar um programa simples e direito de língua inglesa para pessoas que precisavam de aprender rapidamente. Até agora, a *SMART Communications, Inc*¹ tem fornecido um conjunto de *Plain English* (PE) para os serviços administrativos e para o comércio. O outro representante era o *Plain English Campaign* (PEC). Na década de 1970, Chrissie Maher propôs ao governo britânico o uso do inglês simples na vida britânica e lançou a PEC. Depois, em 1979, fundou a PEC, uma empresa de edição comercial e formação sediada no Reino Unido. Em 1981, o Governo respondeu à campanha *Plain English Campaign* lançando o projeto de revisão de formulários públicos sob

a direção de *Sir Derek Rayner*. A empresa também está comprometida com o desenvolvimento do PE, empenhada no treino e edição de PE e outros trabalhos relacionados¹.

De acordo com Spaggiari et al. (2005), nas décadas de 1980 e 1990, a segunda geração de LC apareceu no contexto da tendência de globalização económica e necessidade de localização de documentos, com as principais empresas europeias e americanas a projetar a LC para desenvolver e redigir os documentos técnicos, permitindo aos utilizadores estrangeiros entender os manuais com facilidade e precisão e atendendo às necessidades de venda de produtos e serviços para vários países. O produto representativo desse período é o *ASD-STE100 Simplified Technical English*², da Associação Europeia de Indústrias Aeroespaciais³, que continua a ser desenvolvido, com a 8.ª edição disponibilizada em abril de 2021. O objetivo desse trabalho é melhorar a legibilidade de documentos técnicos. Outro exemplo importante da aplicação da LC surgiu em 1998, quando a *Airbus*⁴ iniciou um projeto dedicado à criação de uma LC para utilização industrial, para aprimorar a qualidade dos textos de advertência na cabine de aeronaves *Airbus*, além de fornecer aos utilizadores um meio para facilitar o seu trabalho, respeitando os rigorosos critérios de segurança. “Este projeto foi dividido em três partes, referindo-se a primeira parte aos aspetos terminológicos, a segunda à sintaxe e a terceira às siglas para nomenclatura de computadores” (Spaggiari et al., 2005, p. 107).

A escrita em linguagem controlada permite que os textos disponíveis para o grande público (manuais de instrução) se tornem mais claros e menos equívocos, o que, muito provavelmente, vai resultar num nível de satisfação muito mais elevado por parte do consumidor (Rascu, 2006, p. 108).

¹ “SMART Communications, Inc. is a leading expert in the implementation of Plain English programs for Government and Consumer English”, cf. sítio web de Plain English da empresa, em <https://www.smartny.com/plainenglish.htm> [consultado em 14-7-2023]. ? Cf. The official home of company PEC, <https://www.plainenglish.co.uk> [consultado em 14-7-2023].

² The official home of ASD Simplified Technical English ASD-STE100 (STE) <http://www.asd-ste100.org>. [consultado em 14-7-2023].

³ European Association of Aerospace Industries (AECMA, now ASD <https://www.asd-europe.org>). [consultado em 14-7-2023].

⁴ Airbus is the largest aeronautics and space company in Europe and a worldwide leader, with roughly 180 locations and 12,000 direct suppliers globally. <https://www.airbus.com/en>. [consultado em 14-7-2023].

Chegado o século XXI, a terceira geração de interesse referente à LC surgiu na época de uso generalizado da TA, e a maioria dos produtos e dos estudos de LC enquadra-se nesse contexto.

Machine translation is potentially one of the most interesting computational applications of controlled language. If a controlled language and machine translation system are attuned to each other, machine translation of texts written in that controlled language can be much more efficient and effective, requiring far less – or ideally even no – human intervention (Somers *et al.*, 2003, p. 254).

No estudo de Sharon O'Brien, em 2003, para obter um conjunto de regras de LC recorreu-se a oito sistemas de inglês controlado em uso ativo: *AECMA Simplified English*, *Attempto Controlled English*⁵⁶, *Alcatel's COGRAM*, *IBM's Easy English*, *GM's CASL*, *Océ's Controlled English*, *Sun Microsystem's Controlled English*, *Avaya's Controlled English*, entre os quais apenas o *AECMA Simplified English* foi classificado como uma LC de orientação humana, sendo os restantes classificados como LC para a máquina. Além disso, o conceito de tradução controlada foi proposto no início do século XXI, constituindo um sinal da combinação formal de TA e LC, rematada pelos sistemas de TA, LC e pós-edição.

This triad of use of controlled languages, machine translation systems and postediting processes coupled with the ever-increasing use of translation memory, computer-assisted translation tools, terminology management systems, etc. depict a new working environment for the professional translator that we call *Controlled Translation*. (Torrejón & Rico, 2002, p. 107).

A 7.^a Conferência Europeia para Tradução Automática de Aplicação de Linguagem Controlada foi realizada pela Associação Europeia para Tradução Automática⁷, na Irlanda, em maio de 2003, com o tema “Tradução Contro-

⁵ Cf. *The official home of “Attempto Project”*, <http://attempto.ifi.uzh.ch/site/description> [consultado em 14-7-2023].?

⁶ Cf. *The official home of “The european association for machine translation”*, <http://www.eamt.org> [consultado em 14-7-2023].

⁷ Cf. *The official home of “The european association for machine translation”*, <http://www.eamt.org> [consultado em 14-7-2023].

lada”. Além do inglês controlado, alguns países europeus cuja língua materna não era o inglês também pesquisaram e desenvolveram as suas próprias línguas controladas, tais como *GIFAS Rationalized French*, *Controlled Language Optimized for Uniform Translation* (CLOUT) etc. O CLOUT foi desenvolvido pelo investigador alemão Uwe Muegge e propôs dez regras para LC de TA:

1. *Write sentences that are shorter than 25 words.*
2. *Write sentences that express only one idea.*
3. *Write the same sentence if you want to express the same content.*
4. *Write sentences that are grammatically complete.*
5. *Write sentences that have a simple grammatical structure.*
6. *Write sentences in the active form.*
7. *Write sentences that repeat the noun instead of using a pronoun.*
8. *Write sentences that use articles to identify nouns.*
9. *Write sentences that use words from a general dictionary.*
10. *Write sentences that use only words with correct spelling* (Muegge, 2007, p. 23).

Nos últimos anos, a investigação sobre LC tem sido enriquecida e já não está limitada às áreas de comunicação para os falantes não nativos, para a redação de documentos técnicos e para a TA. Como um campo interdisciplinar, a LC é também largamente utilizada no meio académico, na indústria e no governo e vai chamar cada vez mais a atenção dos investigadores no futuro.

The study of controlled languages is a very dynamic and highly interdisciplinary field, for the most part occupying small niches in the academic, industrial, and governmental worlds. However, adding all these niches together gives us a large body of past and ongoing work. Assuming that people will have to interact even more closely with computers and across language borders in the future, we will see even more work in this area (Tobias Kuhn, 2014, p. 146).

Nos últimos anos, a investigação relevante sobre a LC tem diminuído, mas o seu valor de investigação ainda não deve ser ignorado. A partir dos tópicos de pesquisa relacionados com *Seventh International Workshop on Controlled Natural Language* (CNL 2020/21)⁸, podemos ver o principal conteúdo de pesquisa das LC de hoje.

⁸ Program of Seventh International Workshop on Controlled Natural Language (CNL 2020/21) <https://www.sigcnl.org/cnl2020.html> [consultado em 14-7-2023].

Possible topics for CNL 2020/21 include: CNL for knowledge representation, CNL for query interfaces, CNL for specifications, CNL for business rules, CNL for dialogue systems, CNL for machine translation, CNL for improved understandability of texts, CNL for natural language generation, design of CNLs, CNL applications, CNL evaluation, usability and acceptance of CNL, CNL grammars and lexica, multilingual CNLs, reasoning in CNL, spoken CNL, CNL in the context of the Semantic Web and Linked Open Data, CNL in the government, CNL in industry, CNL use cases, theoretical properties of CNL (Topics of Seventh International Workshop on Controlled Natural Language, 2020/21).

Como se pode ver, surgiu nos últimos anos uma categoria de investigações sobre LC que merece a nossa atenção, verificando-se também um grande entusiasmo e interesse pelo tema por parte de estudiosos de países diferentes. Com esta tese, com o estudo comparativo de MUE português/chinês, espera-se preencher uma lacuna de investigação relativa à LC entre as duas línguas em questão.

3. Tipos de Linguagens Controladas

De acordo com os objetivos e objetos orientados da LC, podemos dividi-la em dois tipos: “Linguagem Controlada Orientada para o Homem (HOLC)”, que serve para melhorar a legibilidade para leitores humanos, e “Linguagem Controlada Orientada para a Máquina (MOLC)”, que visa melhorar a traduzibilidade para a TA. Huijsen (1998) estabeleceu a distinção entre linguagens orientadas “para o ser humano” e “para o computador”:

a primeira visa esbater problemas de compreensão que a linguagem natural pode originar ao utilizador humano, enquanto a segunda visa eliminar quanto possível os obstáculos que um texto em linguagem natural pode colocar aquando do seu processamento computacional, em particular na tradução automática (Gomes, 2010, p. 3).

A primeira e segunda gerações de LC pertencem principalmente à HOLC, que teve uma aplicação particularmente importante no âmbito do co-

mércio internacional (*Basic English*⁹) e na indústria, especialmente na criação de manuais técnicos (*ASD, Simplified Technical English*¹⁰) e na escrita de avisos sobre o uso de certas máquinas (*Airbus Warning Language*¹¹). Na terceira geração, a tendência evidenciada nos estudos efetuados e as aplicações resultantes são de MOCL. Um dos objetivos das MOCL é tornar o texto “compreensível” e por isso processável por um sistema de TA, seguindo os “critérios de traduzibilidade” (Reuther, 2003). As aplicações das MOCL são múltiplas e este tipo de LC em particular pode ser aplicado à redação de documentos traduzíveis através de um sistema de TA (CTE¹²), na representação e aquisição do conhecimento (ACE¹³, PENG¹⁴, CPL¹⁵) e na construção de redes semânticas (ACE View¹⁶, *Rabbit Lite Natural Language*) (Buchicchio, 2017).

De acordo com Clark et al. (2009), é possível reconhecer duas escolas diferentes no que diz respeito à abordagem utilizada na construção de uma LC: a naturalista e a formalista.

As HOLC seguem a abordagem naturalista, de mais fácil compreensão e utilização por um ser humano e, mesmo controladas, ficam próximas da linguagem natural. As MOCL seguem a abordagem formalista, porque utilizam um único sentido e uma única interpretação aceitável, com uma consequente melhor “compreensão” e utilização por uma máquina (Buchicchio, 2017, p. 6).

É possível encontrar alguns pontos de divergência no que diz respeito à forma como as regras são escritas para HOLC e para MOLC.

⁹ *British American Scientific International Commercial*. Linguagem controlada criada por Charles Key Ogden em 1930 que visa à simplificação da língua inglesa para a comunicação.

¹⁰ ASD STE-100, *Simplified Technical English*. Linguagem controlada para a produção de documentação na indústria aeroespacial, variante simplificada do inglês.

¹¹ Linguagem controlada desenvolvida em 1998 pela *Airbus* para a criação de documentação técnica.

¹² CTE, *Caterpillar Technical English*. Linguagem controlada criada para o sistema de tradução automática KANT e desenvolvida pela Mellon Carneige University junto com a empresa *Caterpillar Inc*.

¹³ ACE, *Attempto Controlled English*. Linguagem controlada para a representação do conhecimento, desenvolvida pela Universidade de Zurique.

¹⁴ PENG, *Processable English*. Linguagem controlada para a representação do conhecimento.

¹⁵ CPL, *Computer Processable English*. Linguagem controlada para a representação do conhecimento, desenvolvida pela *Boeing Research Technology*.

¹⁶ ACE View, *Attempto Controlled English View*.



Naturalist approaches, which view controlled languages as sets of restrictions on the existing structures and lexicon of a given natural language, stating which structures and lexical items are not to be used; and Formalist approaches, which view controlled languages as sets of vocabulary and rules to form utterances in a given natural language, determining the lexicon allowed as well as the syntactic and interpretation rules allowed (Marrafa et al., 2012, p. 153).

No caso das linguagens orientadas para os humanos, as regras podem ser computacionalmente intratáveis e intencionalmente vagas, enquanto no caso das linguagens orientadas para as máquinas, devem ser precisas e computacionalmente tratáveis (Buchicchio, 2017, p. 6).

Em comparação com a HOLC, a MOLC é normalmente de mais difícil utilização por parte do redator e, muitas vezes, pressupõe mais regras e depende de ferramentas sofisticadas para o seu uso, uma vez que para a MOLC existem regras detalhadas sobre terminologia, combinação de vocábulos, comprimento de frase, parágrafo, usos de pontuação etc. Antes de usar uma LC, os redatores de documentos técnicos geralmente precisam de receber formação especial para aprender e memorizar as regras. A experiência mostra que escrever em uma LC não é fácil (Redmond, 1999).

When controlled language is designed for a machine translation system, the constraints on language may be stricter than in a controlled language designed just for authoring. That is because one of the important goals of utilizing controlled language for machine translation is to reduce ambiguity. Because of that, we tend to focus on disambiguation of input sentences when we develop a controlled language. However, usability and author productivity are equally important when deploying a controlled language (Mitamura, 1999, p. 50).

Para facilitar o uso da MOLC para a escrita, os pesquisadores propuseram desenvolver um *software* de verificação de LC (verificador automático LC). O verificador automático LC pode ser usado como um programa de *plug-in* para outros sistemas de escrita (Jeffrey Allen, 1999).

In order to deploy controlled English for production authoring of technical text, an on-line system can be created for interactive checking of texts. This ensures that texts conform to the desired vocabulary and grammar constraints (Mitamura, 1999, p. 48).

Desde os finais do século XX, estudiosos como Mitamura, Nyberg, Kamprath, Almqvist, Hein, entre outros, tentaram investigar e desenvolver os “*controlled language checkers*”, ou seja, os *software* desenvolvidos, tais com *KANT System*¹⁷ e *Scania Checker*¹⁸. As grandes empresas de tecnologia também desenvolveram os verificadores automáticos de LC, entre os quais os representativos são, por exemplo, *SMART MAXit Checker*¹⁹, produzido pela *Smart Communications, Inc.*, e o *software* da *Boeing Simplified English (Boeing-SE Checker)*²⁰ feito pela empresa *Boeing*.

As HOCL e as MOLC têm, como é óbvio, algumas características em comum, como por exemplo a limitação do comprimento das frases e imposição do uso de determinadas estruturas frásicas. “Embora a LC não tenha uma expressão tão rica e colorida como a linguagem natural, possui um grau considerável de capacidade expressiva” (Yuan, 2003, p. 77). A HOCL, apesar de ser controlada, é claramente uma linguagem normativa sempre correta na linguagem natural. A MOLC também deve ser uma linguagem normativa e padronizada, isto é, embora a linguagem seja adequada pela TA, depois de

¹⁷ “The KANT System (Knowledge-based, Accurate Natural-language Translation) has been primarily targeted towards the translation of technical documents written in controlled language. KANT has been developed for 8 multilingual translations of heavy equipment documentation, and is currently in production use for French and Spanish translations. (Mitamura, et al., 1991; Nyberg and Mitamura, 1996).” Cf. Mitamura (1999). *Controlled language for multilingual machine translation*. In *MTSummit* (pp. 46-54).

¹⁸ “Scania Checker, a web-based language checker for Swedish automotive service literature. In order to ensure translation consistency and quality, technical writers at Technical Information use it to check grammar and vocabulary in the source document before it is being translated.” Cf. Almqvist & Hein (2000). *A language checker of controlled language and its integration in a documentation and translation workflow*. In *Proceedings of Translating and the Computer 22*.

¹⁹ “The MAXit Checker is a plug-in to Microsoft Word, Adobe FrameMaker 2020, Visio, PTC Arbortext Editor and other publishing tools. The program uses custom dictionaries and 19,000 AI/NLP rules to simplify technical instructions. See *Simplified Technical English (STE for aerospace)*, *Controlled English* and *Plain English (eGovernment)*”, cf. <http://www.smartny.com> [consultado em 14-7-2023].

²⁰ “The Boeing Simplified English Checker (BSEC) helps technical writers check their documents for compliance with ASD (AeroSpace and Defence Industries Association of Europe) *Simplified Technical English*, a writing standard for aerospace maintenance documentation.”, cf. <https://www.boeing.com/company/key-orgs/licensing/simplified-english-checker.page> [consultado em 14-7-2023].

ser passada para a LC e antes de ser aplicada à TA, os aspetos da linguagem, tais como combinação de termos, comprimento de frase, parágrafo, entre outros, têm de respeitar regras linguísticas. Não podemos usar elementos que contenham erros na linguagem natural visando obter melhores resultados na TA. Seja HOCL ou MOLC, o seu objetivo final é tornar o texto mais fácil de entender. “*It is often difficult to qualify a controlled language as either human-oriented or machine-oriented, since often simplification works both ways*” (Huijsen, 1998, p. 2).

Em certas circunstâncias, a LC pode, por um lado, facilitar a compreensão dos documentos técnicos pelos utilizadores, aprimorando a padronização de documentos técnicos e garantindo a consistência da terminologia e, por outro lado, tornar os textos traduzíveis para TA, melhorando a qualidade da tradução e reduzindo os custos, pelo que podemos dizer que é “matar dois coelhos de uma cajadada” (Yuan, 2003, p. 78).

Esta tese centra-se na LC entre o português e o chinês, nomeadamente os MUE escritos em LC, analisando principalmente a tradução destes textos técnicos entre português e chinês, para que os tradutores possam entender e traduzir facilmente e com precisão, de modo a melhorar a sua eficiência no trabalho. Portanto, a tese centra-se principalmente na HOLC, tendo em conta a possibilidade de os tradutores recorrerem à tradução assistida por máquina no contexto mais vasto da TA, no seguimento de algumas investigações realizadas sobre MOLC, quando preveem esta funcionalidade além da HOLC. O resultado ideal, no final, é constituir modelos inteiros para as empresas portuguesas e chinesas, fazendo os MUE ideais em português/chinês, que, por um lado, podem ajudar os tradutores de textos técnicos chinês/português a evitar erros e melhorar a qualidade de tradução e, por outro lado, podem ser usados na tradução português/chinês assistida por máquina, reduzindo, assim, os custos e melhorando a eficiência da tradução.

4. Linguagem Controlada e pré-edição

A LC e a pré-edição são duas formas de melhorar a qualidade de *output* da TA e ambas têm o objetivo de eliminar as ambiguidades e melhorar a traduzibilidade. Alguns estudiosos dizem que a LC é um tipo de pré-edição, porque as suas regras gramaticais, lexicais e sintáticas podem ser usadas na pré-edição. “Podemos usar as regras de escrita de linguagem controlada para editar os textos antes da TA, e a pré-edição com base na linguagem controlada pode melhorar a qualidade da TA” (Feng & Gao, 2017, p. 63). Contudo, é evidente que a LC e a pré-edição são dois conceitos distintos e o trabalho estudará principalmente a LC, apesar de algumas conclusões poderem, previsivelmente, ser usadas na pré-edição. Por isso, para evitar a confusão entre os conceitos mencionados, apresentamos a seguinte análise sobre as suas diferenças.

Em 1950, E. Reifler propôs a pré-edição, que é uma ação aplicada ao texto antes de o submeter ao computador, para tradução pela máquina, em que os editores de pré-tradução devem eliminar as ambiguidades dos textos originais. O acadêmico famoso de TA Y. Bar-Hillel, em 1971, apontou que a realização de FAHQT (*Full Automatic High Quality Translation*) era muito difícil, enfatizando ao mesmo tempo a importância da pré-edição. É necessário reescrever o texto em maior conformidade com a expressão da língua de chegada, antes que a TA possa ser executada. Esta ação de reescrita é normalmente conhecida como pré-edição (Nagao, 1989). Note-se que, em primeiro lugar, a pré-edição é feita normalmente com base no texto original; portanto, o texto original e o texto pré-editado serão diferentes, constituindo o texto de LC o texto original. Em segundo lugar, os utilizadores de pré-edição geralmente são tradutores individuais e o objetivo é trabalhar com TA visando fazer melhores traduções. No entanto, a LC é mais usada por empresas ou organizações para a conveniência dos seus clientes ou utilizadores, para economizar custos, com a possível melhoria da qualidade de resultados, caso estes textos sejam submetidos a TA.

A pré-edição visa um *output* melhorado na tradução automática, não tem como objetivo, contudo, obter uma tradução perfeita. No âmbito da tradução automática, as linguagens controladas são a forma mais radical de adaptar o texto de partida, uma vez que as alterações efetuadas são

mais abrangentes do que numa fase de pré-edição. Estas podem abranger áreas como o léxico, a sintaxe, ou a semântica. O estilo, que engloba todos os fenómenos que não são passíveis de ser prescritos pelas regras da gramática, também entra no processo de controlo/simplificação (Gomes, 2010, p. 37).

O objetivo da LC é mais abrangente do que apenas melhorar a eficácia da TA, sendo um ajuste mais completo do texto do que a edição pré-tradução, sem se limitar ao ajuste de gramática e vocábulos, podendo incluir também o ajuste de estilo ou estrutura de discurso e assim por diante. O texto controlado pela língua é mais preciso e formal na sua apresentação e mais legível e compreensível para o leitor.

5. Língua Portuguesa Controlada

Os estudos de português controlado iniciaram-se no século XXI. Sandra Fisher Martins é uma das pioneiras na introdução da metodologia *Plain Language* em Portugal. Ela fundou a *Português Claro* (atualmente, apenas *Claro*), uma empresa de consultoria e formação que introduziu o conceito de linguagem clara em 2007, tendo ajudado empresas e instituições portuguesas a comunicar de outra forma²¹. Ao mesmo tempo, estão a ser realizados no Brasil estudos relevantes sobre a simplificação da língua portuguesa. O projeto *PorSimples* foi constituído pelo núcleo interinstitucional de Linguística Computacional da Universidade de São Paulo, entre novembro de 2007 e abril de 2010.

PorSimples: Simplification of Portuguese Texts for Digital Inclusion and Accessibility, The main goal of PorSimples was to develop Natural Language Processing technologies related to Text Adaptation to promote digital inclusion and accessibility for people with low levels of literacy. There are two general different approaches for TA: Text Simplification and Text Elaboration.²²

²¹ Cf. Entrevista com Sandra Fisher-Martins (dezembro 2021). N.º 13 Diagrama. <https://diagrama.ama.gov.pt/w/entrevista-sandra-fisher-martins> [consultado em 14-7-2023].

²² Cf. Introdução a *PorSimples: Simplification of Portuguese Texts for Digital Inclusion and Accessibility* pelo Interinstitutional Center for Computational Linguistics (Núcleo Interinstitucional de Linguística Computacional da Universidade de São Paulo), <http://www.nilc.icmc.usp.br/nilc/index.php/projetos?layout=edit&id=27> [consultado em 14-7-2023].

*WordNet PortControl-CLG-Português Controlado*²³ foi um projeto de estudo desenvolvido pelo Centro de Linguística da Universidade de Lisboa, iniciado em janeiro de 2009 e concluído em janeiro de 2012. De forma subsidiária, recorreu-se ao *WordNet.PT*²⁴. Os resultados tangíveis foram notáveis: um Português Controlado, designado CLG – Português Controlado; uma aplicação computacional que supre as lacunas do conhecimento linguístico do utilizador, designada *WordNet PortControl Checker*.

WordNet PortControl constitutes the first system to control European Portuguese for machine translation and for Portuguese teaching/learning, involving the design of a CNL and of checking tools. It covers basic grammar phenomena such as order of constituents, null constituents, ellipsis, use of pronouns, proper noun identification, auxiliary verb (tense and aspect) system (Marrafa *et al.*, 2012, p. 152).

As regras gerais de linguagem controlada criadas pelo CLG - Grupo de Computação do Conhecimento Léxico-Gramatical – do Centro de Linguística da Universidade de Lisboa para o par linguístico português-inglês, serviram também como base para a criação das regras específicas igualmente propostas neste trabalho.

6. Língua Chinesa Controlada

As palavras chinesas não têm flexões e as classes de palavras são multifuncionais. Por exemplo, os substantivos chineses, além de desempenharem as funções de sujeito e objeto, também podem ser usados como atributos e predicados. Os adjetivos chineses são usados principalmente como atributos, mas muitas vezes também como predicados e complementos. Estas caracterís-

²³ Cf. *WordNet PortControl – CLG-Português Controlado*, <http://clul.ulisboa.pt/projeto/wordnet-portcontrol-clg-portugues-controlado> [consultado em 14-7-2023].

²⁴ A *WordNet.PT* (Marrafa, 2001, 2002) é uma base de dados de conhecimento lexical do português, que está a ser desenvolvida no quadro da *EuroWordNet* (Vossen, 1999), uma base de dados multilingue de larga escala que inclui wordnets de várias línguas europeias. A *EuroWordNet* segue as linhas gerais da *WordNet* de Princeton, que constitui a primeira base de dados de conhecimento linguístico em que o significado lexical é representado através de uma rede de relações lexicais e conceituais, sendo o significado de cada unidade lexical derivado da sua posição na rede (Marrafa *et al.*, 2006).

ticas determinam as dificuldades de análise do chinês, até certo ponto. A LC é, portanto, necessária para reduzir a incidência da ambiguidade na redação de textos formais, especialmente textos de comunicação técnica.

Yu Shiwen propôs o conceito de chinês controlado, em 1995, e abordou a necessidade da pesquisa de chinês controlado. Pouco tempo depois, a *Gramática Controlada do Chinês Moderno (Rascunho)* foi constituída por Jian Sun e outros investigadores do Departamento de Ciência e Tecnologia da Computação da Universidade de Nanjing, em 1996. O estudo constrói regras complementares com certa descrição, incluindo vários exemplos. A gramática controlada inteira apresenta uma descrição de três níveis referentes a regras e disposições suplementares. Zhang Wei apresentou a *Conceção de sistema de escrita com o auxílio de chinês controlado*, em 1998. De 1996 a 2000, o Instituto de Linguística Computacional da Universidade de Pequim e o Centro de Engenharia de Software do Instituto de Tecnologia da Informação de Pequim fizeram, em conjunto, o projeto “Tecnologia de aplicação e desenvolvimento de produtos de chinês controlado”. As pesquisas atuais em chinês controlado são desenvolvidas principalmente na área de MOLC. A plataforma *Pitaya*²⁵, lançada em 2020 pelo Research Center for Language Intelligence of China²⁶, permite à inteligência artificial automatizar a escrita, revisão, retoque ou refinamento de artigos de gêneros controlados.

Os gêneros controlados são os gêneros ou estilos restritos que têm aplicações específicas de utilização e são relativamente fixos em estrutura e formato. A equipa de pesquisa propôs um modelo inteligente de *big data* de expressão chinesa por meio da mineração de dados de textos em massa em estilos ou gêneros controlados, desenvolveu um sistema de escrita inteligente para estilos ou gêneros restritos chineses (Lv, 2020²⁷, p. 80).

Atualmente, com o desenvolvimento de programas como a inteligência artificial para conversação, escrita de artigos, entre outros, a investiga-

²⁵ Cf. Página inicial de *Pitaya*, <https://www.mypitaya.com/en> [consultado em 14-7-2023].

²⁶ Cf. Introdução a Research Center for Language Intelligence of China: <https://yyzn.cnu.edu.cn/yjpt/bjyy-znxtjy/124897.htm> [consultado em 14-7-2023].

²⁷ Cf. Introdução a *Pitaya*, <https://www.mypitaya.com/100080.html> [consultado em 14-7-2023].

ção relacionada com LC chegou a novas áreas, tais como o controlo de estrutura e sintaxe do discurso de gêneros ou estilos especiais.

Conclusão

Com a comunicação cada vez mais estreita entre a China e os países de língua portuguesa, a investigação relacionada com a linguagem controlada entre português e chinês é de grande importância. Desempenhará um papel importante no domínio da tradução automática, no domínio do ensino, no domínio da redação de textos para comunicação técnica e no domínio da inteligência artificial. Por isso, a investigação sobre a linguagem controlada é de grande interesse.

Bibliografia

- Adolphson, E. (1998, May). **Writing instruction and controlled language applications: Panel discussion on standardization.** In Proceedings of the Second International Workshop on Controlled Language Applications (CLAW98) (p. 191).
- Allen, J. (1999). **Adapting the concept of “translation memory” to “authoring memory” for a controlled language writing environment.** In Proceedings of Translating and the Computer 21.
- Almqvist, I., & Hein, A. S. (2000). **A language checker of controlled language and its integration in a documentation and translation workflow.** In Proceedings of Translating and the Computer 22.
- Allen, Jeffrey. (1999). **Different kinds of Controlled Languages.** In TC-Forum magazine, volume 1-99, 4-5.
- Allen, J. (2003). **Controlled Translation: the Integration of Controlled Language and Machine Translation.** Panel session, EAMT/CLAW.
- Crabbe, S. (2017). **Controlling language in industry: Controlled languages for technical documents.** Palgrave Macmillan. DOI 10.1007/978-3-319-52745-1_1
- Cremers, L. (2003). **Controlled Language in an Automated Localization Environment.** Keynote address at EAMT/CLAW.

Kamprath, C., Adolphson, E., Mitamura, T., & Nyberg, E. (1998, May). **Controlled language for multilingual document production: Experience with Caterpillar technical English**. In *Proceedings of the Second International Workshop on Controlled Language Applications* (Vol. 146).

Kathy, B. *et al.* (1999). **GIFAS Rationalized French: A Controlled Language for Aerospace Documentation in French**. *Technical Communication*, Volume 46 (2), 220-229.

Marrafa, P., Amaro, R., Freire, N., & Mendes, S. (2012). **Portuguese controlled language: Coping with ambiguity**. In *Controlled Natural Language: Third International Workshop, CNL 2012, Zurich, Switzerland, August 29-31, 2012*. *Proceedings 3* (pp. 152-166). Springer Berlin Heidelberg.

Mitamura, T. (1999, September). **Controlled language for multilingual machine translation**. In *MTSummit*, 46-54.

Muegge, Uwe. (2007). **Controlled language: The next big thing in translation?** *ClientSide News Magazine*. http://works.bepress.com/uwe_muegge/4.

Nyberg, E. H., & Mitamura, T. (1996, March). **Controlled language and knowledge-based machine translation: Principles and practice** (pp. 74-83). In *Proceedings of the first international workshop on controlled language applications*.

Rascu, E. (2006). **A controlled language approach to text optimization in technical documentation**. *Proc. of KONVENS 2006*, 107-114.

Spaggiari, L., Beaujard, F., & Cannesson, E. (2003). **A controlled language at Airbus** (pp. 107-122). In *EAMT Workshop: Improving MT through other language technology tools: resources and tools for building MT*. <https://doi.org/10.1075/li.28.1.09spa>.

冯全功 & 高琳. (2017). 基于受控语言的译前编辑对机器翻译的影响. *当代外语研究* (02), 63-68+87+110.

胡清平. (2005). 机器翻译中的受控语言. *中国科技翻译* (03), 24-27. <https://doi.org/10.16024/j.cnki.issn1002-0489.2005.03.008>.

胡清平. (2005). 受控语言及其在汉英机器翻译里的应用前景. *国际译联第四届亚洲翻译家论坛论文集*, 134-141.

袁亦宁. (2003). 受控语言: 机器翻译的新途径. *上海科技翻译* (03), 77-80.