Научная статья
УДК 811.111

Лингвистическая теория и вычислительная лингвистика

Аёдижи Акинола Оладеле[1]
[1]Новосибирский государственный педагогический университет, Новосибирск, Россия

Аннотация. Лингвистическая теория и компьютерная лингвистика – это две области, которые тесно переплетены, поскольку обе стремятся понять и смоделировать сложности человеческого языка. В этой статье мы исследуем взаимосвязь между лингвистической теорией и компьютерной лингвистикой и то, как они могут информировать друг друга для улучшения нашего понимания языка. Объединение лингвистической теории и компьютерной лингвистики ведет к еще большему развитию искусственного интеллекта. Обе системы привели к расшифровке, анализу и генерированию человеческого языка. Это было доказано созданием функциональных и поддерживаемых программных приложений, несколькими примерами которых являются видеорегистраторы или виртуальные помощники, службы языкового перевода и Chabots.

Ключевые слова: лингвистическая функция, язык, компьютерная лингвистика, лингвистический подход, языковая функция, искусственный интеллект, информационные технологии, лингвистическая теория, социальные и лингвистические науки

Для цитирования: Оладеле А.А. Лингвистическая теория и вычислительная лингвистика // Актуальные проблемы филологии и методики преподавания иностранных языков. 2024. Т. 18, № 1. С. 75–81.

Original article

Linguistic Theory and Computational Linguistics

Ayodeji A. Oladele[1]
[1]Novosibirsk State Pedagogical University, Novosibirsk, Russia

Abstract. Linguistic theory and computational linguistics are two fields that are closely intertwined, as they both seek to understand and model the complexities of human language. In this article, we will explore the relationship between linguistic theory and computational linguistics, and how they can inform each other to advance our understanding of language. The union of linguistic theory and computational linguistics is leading a greater development in terms of artificial intelligence. Both systems have led to decoding, analyzing and generating human language. This has been proven by the production of functional and enabled software applications of which a few examples are VRs or virtual assistants, language translation services, and Chabots.

Keywords: linguistics function, language, computational linguistics, linguistic approach, language function, AI, Information Technology, linguistic theory, social and linguistic sciences

For citation: Oladele A.A. Linguistic Theory and Computational Linguistics. *Topical issues of philology and methods of foreign language teaching*, 2024, vol. 18, no. 1, pp. 75–81. (In Russ.)

Introduction and Background of the study. The objective of linguistic theory is to understand both the structure and function of a language, on the other hand, computational linguistics aims to develop models and algorithms that analyze and process languages by the use of computers.

Similarly, both are related by the studies of pragmatics, syntax and semantics. However, the approach to these elements is based on different perspectives. It can be stated that linguistic theory enables

the major framework for comprehending functions of language, including the principles that govern language and rule its use. It examines the cognitive processes entailed in not only the production of language but also its comprehension, as well as all factors such as cultural and social that influence language evolution.

Contrarily, computational linguistics makes use of computational and mathematical means to process and analyze data from natural language. This science device software program solves information retrieval, machine translation, and speech recognition by using systems based on linguistic theories. Both linguistic theory and computational linguistics are bound by mutual benefits. This is because Linguistic theory offers a theoretical basis for the computational models of language, while computational linguistics provides the computational tools and empirical data needed to test and redefine linguistic theories.

Essential Requirements for Computational Linguistics. In the process of applying computational linguistics to linguistic theory, varying objectives are always a challenge. Computational linguists focus on creating a system that can modify speech instances and subsequently written or spoken communication to achieve predefined outcomes. In contrast, linguistic theories focus on a language's overall structure, considering individual performance as a result of the interplay between linguistic competence and numerous undetermined factors. Consequently, computational linguists must devise pragmatic solutions to address their specific challenges. Let's explore the theoretical framework suitable for computational linguists. Differentiating formal theories, reliant on formal properties, from functional theories, emphasizing meaning conveyance, offers valuable insights. Functional theories, providing direct access to meaning, align closely with computational linguistics' objectives.

However, most lack the precision required for computer environments. T. Winograd attempted to employ one such theory, requiring substantial reformulation of systemic grammar [9]. Consequently, computational linguistics predominantly relies on formal language theories, beginning with distributional linguistics. However, a sound formal theory should not only enable discourse manipulation but also possess interesting functional attributes, like semantic invariance or logically oriented semantic relations. Various theories exist, notably generative semantics and Lamb's stratificational grammar, proposed for computational language work by R. Binnick and S. Lamb respectively [8]. Despite their significant differences, both theories share common features. Primarily, they function as language theories, aiming to explain the well-formedness and systematic relationships within sentences of a language. As a consequence, they establish systematic, deterministic relations between expression and content. In generative semantics, this relationship flows from content to expression, often resulting in ambiguous interpretations due to the 'undoing' of transformations.

Conversely, stratificational grammar lacks orientation, presenting the understanding of a sentence as the activation of a static network from the expression side, forming a complex of activated lines representing the sentence's meaning at the content end. Notably, both heavily rely on well-formedness, assuming any imperfect speech act to be halted during encoding or decoding, which might be a norm in interactions limited to well-formed speech expressions but unnatural in human communication where well-formedness is not imperative.

Therefore, while generative semantics or stratificational grammar might be artificial simplifications of human language for computational purposes, it's acknowledged that a single sentence can have varied meanings in different contexts, making it unrealistic to seek a purely linguistic mapping between content and expression.

Adopting a mediate theory of language leads to a significant implication: it eliminates the possibility of deriving expression's well-formedness from content, and vice versa. Each aspect requires individual specifications. Consequently, this necessitates the development of a separate metalanguage for content, a task initiated, albeit with limited scope, by symbolic logic. This shift also alters the perspective on syntax. In most formal syntax theories, significant focus is placed on selectional restrictions, which essentially mirror semantic well-formedness requirements and are typically addressed within the semantic description. As a result, syntactic description can now be streamlined to the delineation of order structures and strict subcategorization. This enables the disregard of certain syntactic notions, such as transformation, that predominantly characterize selectional invariance.

Exploring a viable device for simplified syntactic description prompts an exploration through the realm of morphology. M. Halle's paper offers insights into the treatment of morphology within generative grammar. Binnick's review within this framework highlights a crucial aspect of Halle's proposition: the differentiation between irregularities in expression and content, a distinction that a mediate theory accomplishes more effectively than a direct one [6]. Conversely, certain regularities in morphology establish a systematic correspondence between expression and content. For instance, in English, the nominalization of a verb is consistently possible; if a 'strong nominalisation' such as 'the arrival of the prime minister' is unavailable, a 'gerund' like 'the second coming' can be utilized. Notably, these regular formations tend to be less acceptable when a lexical possibility exists, hence classified as 'otherwise' solutions. This suggests an extensive utilization of disjunctive ordering in lexicon description, a characteristic inherent in the structure of stratificational theory.

The dictionary plays a critical role, whether in listing irregularities in morphology, accounting for the linkage between expression and content aspects of morphemes, or potentially fulfilling both functions. This repository manifests as a compilation of arbitrary associations between certain content structures and expression structures, a concept traceable back to Saussure. Drawing from the practices of generative grammarians, one might argue that since the dictionary is indispensable, leveraging it for syntax becomes a plausible approach. This perspective echoes Saussure's notion that syntax cannot wholly be part of language, asserting that solely 'stereotyped expressions' form a linguistic toolkit akin to recurrent syntagmatic patterns of morpheme classes, leaving the rest to the speaker's discretion.

This viewpoint carries two intriguing implications. Firstly, from a computational perspective, any device necessary for handling the dictionary could serve the purpose of managing syntagms. Secondly, from a theoretical standpoint, a theory addressing syntactic ill-formedness could naturally be formulated, drawing an analogy with morphology. To illustrate, in Halle's paper, 'arrivation' is considered well-formed despite not occurring, owing to the specific attributes of '-ation' and 'arrive'. Morphological well-formedness can be articulated in terms of contiguity relationships. Following this model, it

becomes plausible to express syntactic conditions in similar terms of contiguity relationships. An objection frequently raised against merging morphology and syntax revolves around the notion that the level of recursiveness found in syntactic structures lacks resemblance to that in morphology [3].

Harris, in his foundational presentation of the string structure of English, proposed a solution based on the observation that only a few types of syntagms in English syntax exhibit recursion – denoted by Harris. By employing a finite state diagram, English order structures can be depicted, wherein specific transitions are associated with one of these three symbols, triggering a recursive invocation of the corresponding syntagm's description.

A similar scheme may not appear entirely novel to stratificationalists, who may argue that their theory distinguishes between regularities and idiosyncrasies in description and use. They may also contend that what is to a dictionary is simply the set of static relationships represented in a stratificational network. However, there are two key differences to consider. Firstly, it was argued that for a computational device to adequately model linguistic performance, it must be able to handle noise, such as unrecognizable characters in printed form or morphemes of unknown classification. Therefore, message recognition cannot depend on the "activation" of a stratificational network, which would be hindered in both cases.

Work is in progress on combining the capabilities of vigilant memory with the efficiency of conventional dictionary lookup procedures. Importantly, the output of a vigilant memory is a decision on the identity of a form given a possibly faulty input. This output can then be input to another vigilant memory, which will recognize other forms based on the previous ones (e.g. words or syntagms in terms of morphemes) [7]. The proposal here is different from stratificational grammar in an important way. Stratification grammar is a direct, structural theory that does not offer a way to represent an isolated construct of expression or content independently of the general network of relationships describing the overall structure of the language. In contrast, we propose two sets of well-formedness characterizations, one for expression and one for content. The link between the two is seen as a collection of procedures called recognizable forms of expression, which build forms of content according to the well-formed schemas of content. In other words, units of expression (morphemes, syntagms, sentences, etc.) do not have, carry, or correspond to a particular meaning, but induce certain computations whose result is some meaning structure. This view, which Winograd also seems to hold, is at the center of Integrative Semantics [5]

Integrative Semantics. There are several kinds of semantic processes. The simplest kind is the object of logical proof theory, which involves manipulating semantic forms to obtain other systematically related forms. However, our cognitive activity involves more than just manipulating abstract forms. There are two other kinds of semantic processes, also present in Winograd's system. The interpretation procedure consists of combining abstract forms, making them correspond with a portion of some general concept, while the evaluation process concludes about the appropriate response in a specific case. It's vital to notice that these processes are not only about speech activity, they are always involved in our conscious lives. Speech can be considered as an ideal method of influencing cognitive activity, however, it sometimes may not be a success, as in a case of a hearer failing to understand or misunderstand.

Considering human language, an expression can be constrained through conditions well-formed by the limit of possible use and arrangement of morphemes. Based on this, an expression is formed into groups such as phrases and sentences, and may not equate to complete structures of content [4]. Whenever a new type of content is to be passed to a hearer, it is usually in a piecemeal way through different successive sentences. The method of constructing a separate structure of content, named integration, is to be directed by specific integration functions. Similarly, direct interpretation has devices such as deictic, they make descriptions definite, mostly with non-linguistic information. Evaluation functions as exist to point the evaluation of an ideal response. These functors do not have an independent meaning of the exact use of conditions and must be explained in the terms of the computational procedures called for in a hearer. The debate for simplicity shows that we treat lexical items by the same procedure.

Modelling Based on The Hearer's Performance. According to this model, the focus is to a listener, and the speaker's need to take into consideration different factors during communication, an example is the listener's status, knowledge, and some other variables. Although computational linguistics tasks mostly involve the simplification of assumptions, human complexity in interactions is not constantly present. However, the formalization of speech understanding is simpler than the production of coherent discourses from the perspective of linguistic theory [1].

A hearer can be said to be a cognitive system that enables manipulation of semantic forms, such as objects, relations, descriptions and also modalities including moods and quantifiers. Manipulating these forms results to semantic forms and judgments of implication, consistency, and contradiction.

It is possible that the cognitive procedures used for building semantic forms in non-linguistic situations are the same as those used for lexical items and integrative functors. This aligns with Whorf's hypothesis. Lexical items and integration functors are seen as subroutine names that trigger building procedures, while integration factors call forth procedures for interrelating parts of semantic forms.

A listener can connect semantic forms to an interpretation universe, which is traditionally referred to as attention. This means that interpretation functions guide the listener's attention and can involve procedures to determine the best referent for a given description. Some interpretation functions also guide the selection of an appropriate reference universe. Similarly, performatives and other devices prompt procedures for evaluation in preparation for an appropriate response [10].

The various procedures prompted by elements of expression in discourse are often not enough on their own to produce all the mentioned results. Part of the speaker's communicative ability lies in omitting much of the necessary information to build, interpret, and evaluate semantic forms, and only selecting what is absolutely essential to guide the listener in the right direction. This is why the term "nudging" was used in the remaining information that the speaker leaves out must be provided by the listener, either through independent manipulation of their semantic forms or by drawing on their store of previous information, known as their "knowledge of the world." [2] If syntax is indeed reducible, in its "order structure" aspect, to a set of patterns recognizable by a device like a Vigilant Memory, then we have an explanation of why syntactic well-formedness is not more destructive. From the listener's perspective, the only thing that matters is being able to recognize a pattern present in

their syntactic dictionary, which will then prompt a building or integrating procedure. Incorrect syntax only becomes destructive when the error-correcting capabilities of the vigilant memory are overwhelmed. However, it is a constant possibility even with the occurrence of recognition, to detect syntactic informedness by the workings of a memory.

Conclusion. The combination of linguistic theory and computational linguistics is important for advancement of understanding human language. The integration of insights in these fields, has helped researchers to develop more accurate and better language models that are being applied for real-world problems. As technology keeps advancing, the union of linguistic theory and computational linguistics will be pivotal in developing advanced computer programs for both natural language processing and general understanding. Also, the joint efforts of linguistic theory and computational linguistics will continue to add to the advancements of tools for text summarization, grammar correction, and translation. These programs have different uses such as sale and marketing, data analysis, and customer service. This is because there is a need for understanding and processing human language in these fields. Lastly, linguistic theory and computational linguistics are mainstream and major factors for the understanding of human languages and the development of technology in the spare of linguistics. The continuous evolution of these fields in the right direction will be important for solving future challenges and create new opportunities in both study and application use of language.

Список источников

1. *Alishahi A., Stevenson S.* A computational usage-based model for learning general properties of semantic roles // Proceedings of the 2nd European Cognitive Science Conference. 1st Edition 2007. P. 425–430.

2. *Boersma P., Hayes B.* Empirical tests of the gradual learning algorithm // Linguistic Inquiry. 2001. Vol. 32, Issue 1, P. 45–86.

3. *Crabbé B.* Grammatical development with XMG // Logical Aspects of Computational Linguistics. 2005. № 3492. P. 84–100.

4. *Daumé III.* H., *Campbell L. A* Bayesian model for discovering typological implications // Conference of the Association for Computational Linguistics (ACL). 2007. Vol. 1. P. 65–72.

5. *Payet J.P.* Prérequis pour une théorie semantique // Cahier de linguistique. 1973. № 2. P. 14–19.

6. *Payet J.P.* Computational linguistics and linguistic theory // Proceedings of the 5th conference on Computational linguistics. 1973. Vol. 2. P. 357–366.

7. *Douglas R., Jurafsky D.* How verb subcategorization frequencies are affected by corpus choice. // Proceedings of the 17th International Conference on Computational Linguistics. 1998. Vol. 2. P. 1122–1128.

8. *Lamb S.* Readings in Stratificational Linguistics / Edited by Adam Makkai and David G. Lockwood. University of Alabama Press 1973. 320 p.

9. *Winograd T.* Understanding Natural Language; Edition 6, illustrated, reprint; Publisher Academic Press, 1972. 191 p.

10. *Woods W.A.* Transition networks for natural language analysis // Communications of the ACM. 1970. № XIII. P. 591–606.

Информация об авторе

А.А. Оладеле – ассистент кафедры лингвистики и теории перевода, Новосибирский государственный педагогический университет, joyfullayo@gmail.com

Information about the author

A.A. Oladele – Assistant of the Department of Linguistics and Translation Theory, Novosibirsk State Pedagogical University, joyfullayo@gmail.com

Статья поступила в редакцию 05.10.2023; одобрена после рецензирования 15.10.2023; принята к публикации 27.10.2023.

The article was submitted 05.10.2023; approved after reviewing 15.10.2023; accepted for publication 27.10.2023.