Computational Linguistics Volume 35, Number 4
2009 Association for Computational Linguistics

自然语言工程的支撑科学是什么？

Shuly Wintner
University of Haifa

我最近听到的最发人深省的见解之一，来自 Lori Levin 在 EACL2009 的语言学与计算语言学互动专题讨论上提出的建议。他建议我们在 ACL 里设置一个语言学专委会。最初的尴尬之下，我发觉这个想法很怪异：这岂不是有点像美国儿科学会妄想设立“医学专委会”或者“儿童专委会”一样滑稽吗？然而接下来一想，才意识到这个想法的合理性：本质上，当代的自然语言工程里，语言学整体上是缺位的！在下文中，我想呼吁语言学回归计算语言学。

二十多年来，我们完成了计算语言学研究范式的整体转型。由于语言学知识在数据规模扩张到真实世界的需求后无法应用带来的沮丧，以及更深刻地，由于形式语言占统治地位的理论带来的沮丧，我们转向了语料库，转向了把语言的使用作为我们知识的（潜在）源泉。与方法论的转型相伴生的，是我们整个行当的目标的微妙变化。在二十年前，一个计算语言学家或许既对开发 NLP 应用感兴趣，也对语言学过程的形式化以及推理感兴趣。现如今，兴趣只剩下前者了。粗粗看一看我们领域主要会议上的文章，绝大多数都是工程型的，讨论的都是实际问题的工程解决方案，几乎不再有人讨论基础性的语言学问题。

当然并不是说工程性的工作有什么错。每个技术院校都设有工程类的系，其领域之广泛已经涵盖诸如化工、机械工程、航空工程，生物医学工程等等。没有理由说不该有一个自然语言工程的学科。但是，就大多数已经设立的学科来看，工程类的系所进行的研究，都是在科学领域里一个非常成熟的分支的指导下进行的。化学工程师研究化学，电气工程师研究物理学，航空工程师研究动力学，生物医学工程师研究生物学、生理学和生命科学，如此等等。

工程的成功，也有选择使用“正确的”数学的功劳。一些科学领域理论上的发展，最典型的如物理学，是与数学的发展比翼齐飞的。没有数学的这些基础性的支撑，物理学就不可能胜任对自然现象的准确刻画。例如，（偏）微分方程的发展，就与物理学上最了不起的发现并驾齐驱，这个数学分支后来又应用于化学、电力和经济学等科学领域。

那么，什么才是给自然语言工程作后盾的科学分支呢？什么才是我们建立应用时所依赖的理论基础支撑呢？当我们对人类语言进行推理的时候，什么样的数学才是必要的呢？

考察一下面向数据的革命以来在自然语言工程领域的重大成果吧。比如宾州树库，1992 年第一版本问世以来，它的标注体系被用来对众多的词类和分析器进行训

练。这套标注体系的背后是什么理论？在什么意义上这个标注体系是“正确的”？会不会有其他某个标注体系也是同样好的？我们凭借什么准则对这样一套资源的质量进行评估？这准则又该嵌入科学的哪一个分支？再看半个世纪以来自然语言处理领域的“皇冠上的明珠” ——机器翻译。我们现在的统计机器翻译系统已经做得足够在很广泛的一类应用场合下使用了，Google 推出了超过 40 种语言两两之间的免费互译服务，你可以在阿尔巴尼亚语和越南语之间互相翻译了。这恐怕算是我们这个领域最伟大的成就了吧，但是它到底基于哪个学科？受到哪个理论的支撑？类似的例子举不胜举。词汇歧义消解、随机句法分析、文本分类、问答、语义角色标记、语音识别、本体开发，随便什么你感兴趣的应用，都可以同样追问：基于什么学科？受到哪个理论的支撑？它的理论支点在哪里？在过去，扮演这个角色的曾经是语言学。词形分析器就是词形学和语音学研究者们的智慧集大成之后的反映。首个句法分析算法是在当时的句法理论的指导下实现的。对话系统吸取了语义学和语篇理论的研究成果。为什么后来不是这样了呢？一个明显的原因就是，依赖清晰的语言学知识作为基础的应用系统在规模扩大时表现不佳。另一个更微妙的原因涉及到科研经费的投入方式：经费投入机构（主要在美国）主要由短期实用目标所驱动，在基础性的研究方面缺乏足够的耐心。计算机科学的其他一些领域从基础性的、理论性的研究转向工程应用，也是相同道理。

但是还有更深层的原因。语言学作为一个学科，正在走向迷失——它只关注句法，而且又以英语的句法为主；它的理论变得如此晦涩，如此变态，如此自恋，以至于其他领域的研究者事实上无法跨学科参与进来。用 Evans 和 Levinson（待出版）的话说，“相关的文献对于圈外人士来说是难以逾越的屏障”，或者用 Tomasello（1995，136 页）的话说，语言学理论“用特殊术语来描述语言，导致我们很难把它与心理学的其他研究领域中关于认知的成果相联系。”至少对于计算机实现而言，是这个原因。因此，我们对语言学感到沮丧，我们在沮丧中彻底放弃了它，剩下的就只有统计学和概率论。但是，计算语言学肯定不是应用统计学的一个分支。假如真是应用统计学的话，那自然语言处理系统就和其他非语言的字符串处理系统，比如 DNA 序列、乐谱、棋谱等非语言学的处理系统没有什么区别了。我们的系统所处理的字符串肯定有某种唯一的特性，有某种可以从理论角度加以概括、在科学意义上加以研究的东西。决定我们的系统的特殊性的，正是在于它处理的是自然语言，而能给我们以指导的唯一的科学领域就是语言学。实际上，在语言学的世界里新东西越多，计算语言学能从中受益的就越多。生成语言学与理论语言学之间一度咬得很紧，但是已经松下来很长时间了。有几个很棒的方向，如能从形式上、数学上、计算上加以研究，相信会能大有裨益。让我举几个例子。

最突出的例子是心理语言学。一些具有非凡的语言天赋、同时又有良好的认知科学素养的研究者，把语言放在更具共性的认知背景下，将一些更有前途的科学方法论引入了语言学研究。让我们看看《儿童》（Mac Whinney 2000），一个收集了数十个研究者在二十多年里覆盖 25 种语言的、描述儿童与其监护人的语言互动的大型语料库。这些语料中的大多数都在构词形态上做了标记（有些甚至做了句法标记），有 1500 多篇研究及评估语言发展当中各种问题的科学论文中使用了这些语料。这些

语料为由计算驱动的语言学研究提供了难以估量的资源：心理语言学、理论语言学、认知语言学以及计算语言学本身都在其列。

心理语言学不仅提供资源和方法论，也提供值得我们计算语言学者去审视、评估和发挥的具体理论。别不多提，像 Mac Whinney（1987,1998,1999,2004a,2004b）、Tomasello(1998,2003,2006)以及 Bybee（2001,2006,2007）等工作就产生了得到实验支持的一些令人兴奋的理论。有趣的是，这些语言学理论暗示了一种我们所需的可以对自然语言进行理解和推理的数学：这些理论强调语言使用而不是抽象的“语言能力”，将我们关注的焦点从形式语法和逻辑引导到统计和概率上来。最重要的是，这些理论在原则上是可证伪的（Tomasello 2004），我们能结合我们自己的数学、计算的方法论对之进行改进吗？我们能用之设计一个更好的系统吗？

令人兴奋的语言学研究成果，植根于在哲学上、生物上、认知上、计算上具有更广泛的洞察力的那种，同样还来自语言进化论（Christiansen 与 Kirby 2003）、历史语言学（Warnow 1997; Nakhleh, Ringe, 及 Warnow 2005），以及来自更加传统的核心领域——词形学以及句法学（Goldberg 1995; Prince 与 Smolensky 1997; Pullum 2007）。我们不仅必须了解这类工作，而且还必须走的更远。我们必须更多地了解能对我们的工程性工作有所改进的语言学研究成果，还必须直接参与到这类研究中来。我们在形式化的数学和计算方面的素养对于生命科学和人文科学来说是无价之宝，可以在传统的研究思路一筹莫展之处助上一臂之力。我们可以为语言学的几乎所有分支带来新见解、新观点。计算语言学本质上可以成为语言学的一个子领域。实际上，在我们的主流会议上，已经可以看到这方面的一些初步的尝试。例如，Ellison 和 Kirby（2006,273 页）提出了一个计算的方法，来构造一个通用的语言分类体系，他们声称该体系可以“更好地融合语言学和心理语言学中的当代思想”。Daum’e III 和 Campbell (2007)使用了一种计算方法论，用于从《世界语言结构图谱》（一个涉及 2000 种语言、诸多属性的大型数据库）中发现有类型学意义的语言共性。在这些例子中，新的计算技术用于辅助语言学家开展他们一直感兴趣的传统类型的研究，是相当成功的案例。

不幸的是。这些例子实在还是太少，而且彼此相距太远。但是它们说明，计算语言学一旦被语言学武装起来，可以达到什么样的成就。我不好说这样的理论一定能够给工程项目许诺什么样的改进，但是自然科学的所有领域都是如此。语言工程也不会是例外。更重要的是，我们的圈子有能力作为当代语言学研究的一支主力后援部队。我们当中的很多人被吸引到这个领域来，是出于我们对语言的热爱。我们现在可以追随我们的内心，回去探索语言的奥秘和精美，而不是……这么说吧……同时也为了满足我们实际需要去利用语言。让我们成为语言学的一份子。不然的话，说不准我们当中的某些人真的要在 ACL SIG 里成为难民了。

致谢

我要感谢 Robert Dale 邀请我为这个栏目撰稿，并感谢他为此所作的准备。我也非

常感谢 Nissim Francez 和 Nurit Melnik 对本文初稿所提的建议。当然这里所表达的观点， 包括其中的错误和误解，都属于我本人。

参考文献

Bybee, Joan. 2001. *Phonology and Language Use*. Cambridge University Press, Cambridge. Bybee, Joan. 2006. From usage to grammar: The mind’s response to repetition. *Language*, 82(4):711–733.

Bybee, Joan. 2007. *Frequency of Use and the Organization of Language*. Oxford University Press, Oxford.

Christiansen, Morten H. and Simon Kirby, editors. 2003. *Language Evolution*. Studies in the Evolution of Language. Oxford University Press, Oxford.

Daum’e III, Hal and Lyle Campbell. 2007. A Bayesian model for discovering typological implications. In *Conference of the Association for Computational Linguistics (ACL)*, pages 65–72, Prague.

Ellison, T. Mark and Simon Kirby. 2006. Measuring language divergence by intra-lexical comparison. In *ACL-44: Proceedings of the 21st International Conference on Computational Linguistics and the 44th Annual Meeting of the Association for Computational Linguistics*, pages 273–280, Morristown, NJ.

Evans, Nicholas and Stephen Levinson. In press. The myth of language universals: Language diversity and its importance for cognitive science. *Behavioral and Brain Sciences*.

Goldberg, Adele. 1995. *Constructions. A Construction Grammar Approach to Argument Structure*. University of Chicago Press, Chicago, IL. MacWhinney, Brian, editor. 1987. *Mechanisms of Language Acquisition*. Lawrence Erlbaum Associates, Hillsdale, NJ.

MacWhinney, Brian. 1998. Models of the emergence of language. *Annual Review of Psychology*, 49:199–227.

MacWhinney, Brian, editor. 1999. *The Emergence of Language*. Carnegie Mellon Symposia on Cognition. Lawrence Erlbaum Associates, Mahwah, NJ.

MacWhinney, Brian. 2000. *The CHILDES Project: Tools for Analyzing Talk*. Lawrence Erlbaum Associates, Mahwah, NJ, third edition.

MacWhinney, Brian. 2004a. A multiple process solution to the logical problem of language acquisition. *Journal of Child Language*, 31:883–914.

MacWhinney, Brian. 2004b. A unified model of language acquisition. In J. Kroll and A. De Groot, editors, *Handbook of Bilingualism: Psycholinguistic Approaches*. Oxford University Press, Oxford, pages 49–67.

Nakhleh, Luay, Don Ringe, and Tandy Warnow. 2005. Perfect phylogenetic networks: A new methodology for reconstructing the evolutionary history of natural languages. *Language*, 81(2):382–420.

Prince, Alan and Paul Smolensky. 1997. Optimality: From neural networks to universal grammar. *Science*, 275:1604–1610.

Pullum, Geoffrey K. 2007. The evolution of model-theoretic frameworks in linguistics. In *Proceedings of the ESSLLI 2007 Workshop on Model-Theoretic Syntax*, pages 1–10, Dublin.

Tomasello, Michael. 1995. Language is not an instinct. *Cognitive Development*, 10:131–156.
Tomasello, Michael. 1998. The return of constructions. *Journal of Child Language*, 25:431–442.
Tomasello, Michael. 2003. *Constructing a Language*. Harvard University Press, Cambridge and London.
Tomasello, Michael. 2004. What kind of evidence could refute the UG hypothesis? *Studies in Language*, 28(3):642–645.
Tomasello, Michael. 2006. Acquiring linguistic constructions. In D. Kuhn and R. Siegler, editors, *Handbook of Child Psychology*. Wiley, NY, pages 255–298.
Warnow, Tandy. 1997. Mathematical approaches to comparative linguistics. *Proceedings of the National Academy of Sciences*, 94(13):6585–6590.

Personal details

Mailing address	Department of Computer Science University of Haifa 31905 Haifa, Israel.
Phone	+972-4-8288180
Fax	+972-4-8249331
E-mail	shuly@cs.haifa.ac.il
Home address	Kadima 5a, 34382 Haifa, Israel
Home phone	+972-4-8364836

Brief bio

Current position
Senior lecturer, Department of Computer Science, University of Haifa, Israel.

1998-2000
Postdoctoral fellow, Institute for Research in Cognitive Science, University of Pennsylvania, Philadelphia.

1997-1998
Postdoctoral fellow, Seminar für Sprachwissenschaft, University of Tübingen, Germany.

1994-1997
D.Sc., Department of Computer Science, Technion, Israel Institute of Technology, Haifa. Advisor: Nissim Francez.