Pragmatics in Learner Corpora

NINA VYATKINA

Pragmatics has been broadly defined as the study of how people behave as social actors using language to accomplish specific communicative goals. Learner corpora are large electronic collections of written or spoken texts produced by learners of a second language (L2) in that language. Thus researchers of pragmatics in learner corpora explore how people learn to achieve communicative goals using an L2 by analyzing a body of texts produced by these learners in their target language.

Both pragmatics and learner corpus analysis are very young subdisciplines of second language studies. Interlanguage pragmatics, or L2 pragmatics, which investigates how L2 learners develop the ability to understand and perform action in a target language, arose as a separate field in the 1990s, whereas learner corpus analysis emerged as a method of research into learner language at the turn of the new millennium. Pragmatics and corpus analysis came together because both disciplines focus on actual language *use* as opposed to abstract qualities of language as a system. The main appeal of corpora for researchers into L2 pragmatics lies in the availability of large amounts of naturally occurring data, which gives them a much wider empirical grounding and supplements data elicited by traditionally used instruments such as questionnaires. Furthermore, corpus analysis methods allow automatic search for recurring patterns which can be analyzed using both quantitative and qualitative methods. These methods allow researchers to elicit information about the distribution and frequency of focal pragmatic elements in the target corpus, as well as about their cooccurrence with other elements. This latter technique, also called analysis of collocations, is especially valuable for pragmaticians, who are less concerned with grammatical accuracy of utterances than with their contextual appropriateness.

Initially, pragmaticians became interested in using native speaker (NS) corpora for both research and L2 instruction purposes. Many of these studies contained pedagogical suggestions for using NS corpora as rich repositories of authentic examples for creation of L2 teaching materials or for direct perusal by learners under the teacher's guidance. Around the turn of the century, this data-driven learning methodology was developed to include not only NS corpus data but also learner corpus data. The most important criterion of suitability of a learner corpus for research in pragmatics is the presence of explicit design criteria. As Granger (2002, p. 9) points out, "the usefulness of a learner corpus is directly proportional to the care that has been exerted in controlling and encoding the variables." Most important variables, according to Granger, are learner variables (e.g., learner context, mother tongue, level of proficiency) and task variables (e.g., time limit, use of reference tools, audience/interlocutor). Given the core position of the user and context in pragmatics, careful documentation of these corpus variables is of vital importance, which makes heterogeneous unspecified corpora unsuitable for L2 pragmatics research.

Beyond such metadata mark-up, corpora may be raw or annotated at different levels including part-of-speech tagging, syntactic parsing, discourse-pragmatic annotation, and error annotation of learner texts. Because such annotation is extremely time- and effort-consuming, the majority of available corpora to date remain raw, with a number of notable exceptions. Corpus annotation for pragmatic categories has been especially rare due to the inherent fuzziness and lack of distinct dividing lines between pragmatic categories such as speech acts. For example, an apology may partly overlap with a request rather than be

The Encyclopedia of Applied Linguistics, Edited by Carol A. Chapelle.
 Published 2013 by Blackwell Publishing Ltd.
DOI: 10.1002/9781405198431.wbeal0936

strictly separated from one another in speech. Therefore, an overwhelming majority of studies have been "word-based" rather than "category-based" (Hunston, 2002). For example, Gilquin (2008) explored how EFL learners used discourse markers in their speech. To extract linguistic data for her research, she specified the list of the target features to include such words and phrases as *you know*, *well*, *actually*, *sort of*, and ran an automatic search for these words on her learner corpus. In this bottom-up approach, researchers explore contextual use of specific linguistic features and then make inferences about pragmatic functions expressed by these words in learner texts. By contrast, researchers who are interested in finding out what sets of linguistic features are used by learners to express specific communicative functions apply a top-down approach and perform needs-based tagging of the focal corpora according to their specific research questions prior to the study. Connor, Precht, and Upton (2002) explored rhetorical devices used in an integrated learner/NS corpus of simulated job applications. First, the researchers defined what meaning units termed rhetorical moves (e.g., "identify the source of information," "state desire for consideration," "provide supporting arguments") constituted the genre of application letters and compiled them into a rubric. Next, they manually tagged all corpus texts according to this rubric, and finally performed an automatic search on the focal tags. This analysis allowed the authors to identify similarities and differences in the discourse structure of application letters written by L2 English learners from different cultural backgrounds.

According to the medium, corpora are subdivided into written, spoken, and hybrid or multimedia corpora. Written learner corpora have been primarily explored using contrastive interlanguage analysis, or CIA (see Granger, 2002). This method enables comparison of learners' L2 performance from the learner corpus with the first language (L1) performance of NSs from an NS corpus to discover differences and similarities in the language use of these two populations. Alternatively, researchers can compare different varieties of learner language (e.g., L2 use by more and less proficient learners). Pragmatic phenomena that have been best explored in written learner language include discourse markers, expressions of modality, and the use of formulaic language. The International Corpus of Learner English, or ICLE (http://cecl.fltr.ucl.ac.be/Cecl-Projects/Icle/icle.htm/), has provided material for numerous comparisons of pragmatic feature use by NSs and by advanced learners of English coming from a variety of L1 backgrounds. Aijmer (2002) showed that learners used significantly more modal verbs in their writing than NSs, which contributed to an overuse of spoken patterns inappropriate in a written academic genre. This result has also been corroborated on the material of a different learner corpus with the reversed configuration of languages: advanced L1 English learners overused modal expressions in their L2 German (Maden-Weinberger, 2008). Nesselhauf's (2004) publication, based on the L1 German subcorpus of ICLE, focused on learner use of verb–noun phrases such as "to have a look" or "to make a mistake." She found that even advanced learners frequently used inappropriate verbs in such phrases (e.g., "to do a mistake"). Gilquin, Granger, and Paquot (2007) investigated phraseological patterns used by learners in argumentative essays to express such typical rhetorical functions of English for academic purposes (EAP) as introducing, comparing, contrasting, and summarizing information. The researchers found that, without making grammatical mistakes, learners used some of these patterns inappropriately: for example, they overused intensifiers (e.g., *absolutely*, *of course*) and underused hedges (e.g., *possibly*, *presumably*) in comparison to an NS college writers' group. This research resulted in a pedagogical application: the new corpus-based edition of the Macmillan English dictionary for advanced learners (Rundell, 2007) is interspersed with paragraph-long tips titled "Be careful!" and "Get it right" following the words and phrases habitually misused by ICLE participants. Rhetorical devices typical of EAP were also explored in expert and learner writing on the material of specialized corpora (e.g., Hyland, 2004).

Compilation of a spoken learner corpus presents an additional challenge because audio-recorded data need to be transcribed prior to analysis. In addition, even when transcribed in accordance with uniform conventions, the spoken language is more difficult to analyze with automatic methods because of significant variation in pronunciation of the same words. Despite this fact, spoken corpora have been attracting increasing attention because of the centrality of human interaction in pragmatics. One research direction is represented by Hasselgren (2002) and Gilquin (2008), who focus on Norwegian and French EFL learners, respectively. Both studies explored the use of discourse markers in speech and came to a converging result that increased speech fluency goes along with the use of more discourse markers such as *like*, *I mean*, *you know*, but fewer hesitation markers such as pauses. An example of research based on speech acts is Reinhardt (2007) who compiled and explored a corpus of oral data produced by international teaching assistants (ITAs) in advising and lecturing role plays. He compared the ITA use of directives with the NS comparison baseline elicited from MICASE (the Michigan Corpus of Academic Spoken English; http://quod.lib.umich.edu/m/micase/). He showed that ITAs underused mitigators and pragmatic features expressing interpersonal involvement in comparison to NSs. Reinhardt's study concludes with a pedagogical suggestion for corpus-based teaching. As opposed to word-based studies addressed above, Pavlenko and Driagina's (2007) took a concept-based angle in a comparative study of the use of emotion vocabulary in speech and showed that conceptual differences between English and Russian present difficulties even for advanced learners of these languages. Yet another direction in spoken corpus research is exploration of intonation patterns. Ramirez-Verdugo and Romero-Trillo (2005), for instance, found tone and pitch differences in realization of some speech acts by L1 and L2 speakers of English.

Diachronic corpora that contain data elicited from the same learners over longer periods of time are especially valuable for studying pragmatic development of L2 learners. However, these are rare because collecting longitudinal data is especially difficult. Nevertheless, a number of longitudinal learner corpora have been collected and have triggered innovative studies into L2 pragmatics. Researchers at Georgetown University have collected and investigated a longitudinal corpus of written productions of learners of German, tracking them over the course of three curricular levels. Among other focal features, Byrnes (2009) investigated the development of the grammatical metaphor (primarily expressed in increasing nominalization) in learner writing. The longitudinal approach to data analysis allowed the researcher to demonstrate the nonlinearity and variability of developmental paths taken by individual learners. Another example is a unique longitudinal multimedia corpus that includes audio and video recorded classroom interaction data that have been collected over several years at Portland State University. Using the conversation analysis methodology, Hellermann conducted several case studies using this corpus to track the development of individual learners from beginning to intermediate levels of ESL proficiency. Focusing on specific pragmatic phenomena such as saying "no" in refusals, Hellermann (2009) conducts microanalyses of the talk-in-interaction at multiple points in time and shows how the learner is gradually socialized in the classroom community of practice.

Another invaluable resource that can be accessed on the internet is the French Learner Language Oral Corpora (FLLOC) database that compiles a number of spoken L2 French corpora (http://www.flloc.soton.ac.uk/). Myles, Mitchell, and Hooper (1999) explored formulaic language and creative constructions in learner language with one of these corpora and showed both common developmental trends and dynamic interaction between these features. Finally, a number of longitudinal corpora collected in untutored L2 acquisition settings from adult immigrants have served as the basis for multiple studies of L2 pragmatics (see Skiba, Dittmar, & Bressem, 2008). These are currently being systematized in

accordance with recognized computational standards and published on the internet by the Max Planck Institute in Nijmegen (http://corpus1.mpi.nl/ds/imdi_browser/).

Most recently, learner corpora containing texts produced in so-called hybrid media such as computer-mediated communication (CMC) have proven rich in pragmatic phenomena. Such corpora have an advantage of containing interactive exchanges typical of spoken conversations yet do not require tedious transcription since the data are produced in a written medium. Classroom-based CMC between learners and NSs can be especially conducive to study of L2 pragmatic development as shown by Belz and Vyatkina (2008). In this study, the synchronous (electronic chat) and asynchronous (e-mail) interactions between learners and NSs revolved around a semester-long class project where the American and German participants discussed intercultural topics and wrote collaborative essays that triggered involvement in a variety of speech acts such as invitation, refusal, and apology. Electronic data produced by learners in CMC interactions with their NS interlocutors were automatically saved and then organized into an electronic corpus. Importantly for CIA research, this corpus also included a built-in NS corpus that served as a valid comparison baseline for learner productions since the NS data came as part of the very same interactions, not from an unspecified external corpus. Moreover, the corpus data were used for a concurrent pedagogical intervention in which learners engaged in corpus-based learning comparing their own and NS use of the focal pragmatic features (German modal particles and pronominal adverbs). The study showed that this approach facilitated L2 pragmatic development because students significantly increased their focal feature use and even approximated the NS norm with respect to frequency and appropriateness. However, the study also highlighted persistent differences in collocational patterns used by learners and NSs.

To advance corpus-based research into L2 pragmatics, there is a need for many more studies on publicly available corpora. New corpora also need to be collected, especially oral and multimedia corpora based on a variety of languages collected from learners at different proficiency levels, as well as developmental corpora tracking the same learners over longer periods of time. Given the resources and time that go into pragmatic annotation, sharing annotated corpora is especially desirable. An excellent example is an undertaking by Maynard and Leicher (2007), who set out to manually annotate MICASE. They plan to annotate this corpus for 25 pragmatic features including discourse style and interactivity (e.g., disagreement, request, humor, sarcasm) that will be computer-searchable. This project is especially valuable because MICASE is freely available and comprises both NS and learner speech samples, thus providing a valid NS comparison baseline for performing CIA. When completed, the annotation of this corpus will facilitate investigations that are lacking to date: ones that will go beyond words and phrases and extend into speech acts and communicative functions which may have multiple linguistic realizations.

SEE ALSO: Corpora in the Language-Teaching Classroom; Interlanguage Pragmatics; Learner Corpora; Pragmatics of Second Language Computer-Mediated Communication; Teaching Pragmatics

References

Aijmer, K. (2002). Modality in advanced Swedish learners' written interlanguage. In S. Granger, J. Hung, & S. Petch-Tyson (Eds.), *Computer learner corpora, second language acquisition and foreign language teaching* (pp. 55–76). Amsterdam, Netherlands: John Benjamins.

Belz, J. A., & Vyatkina, N. (2008). The pedagogical mediation of a developmental learner corpus for classroom-based language instruction. *Language Learning and Technology, 12*(3), 33–52.

Byrnes, H. (2009). Emergent L2 German writing ability in a curricular context: A longitudinal study of grammatical metaphor. *Linguistics and Education, 20*(1), 50–66.

Connor, U., Precht, K., & Upton, T. (2002). Business English: learner data from Belgium, Finland and the U.S. In S. Granger, J. Hung, & S. Petch-Tyson (Eds.), *Computer learner corpora, second language acquisition and foreign language teaching* (pp. 175–94). Amsterdam, Netherlands: John Benjamins.

Gilquin, G. (2008). Hesitation markers among EFL learners: Pragmatic deficiency or difference? In J. Romero-Trillo (Ed.), *Pragmatics and corpus linguistics: A mutualistic entente* (pp. 119–49). Berlin, Germany: De Gruyter.

Gilquin, G., Granger, S., & Paquot, M. (2007). Learner corpora: The missing link in EAP pedagogy. *Journal of English for Academic Purposes, 6*(4), 319–35.

Granger, S. (2002). A bird's-eye view of learner corpus research. In S. Granger, J. Hung, & S. Petch-Tyson (Eds.), *Computer learner corpora, second language acquisition and foreign language teaching* (pp. 3–33). Amsterdam, Netherlands: John Benjamins.

Hasselgren, A. (2002). Learner corpora and language testing: Smallwords as markers of learner fluency. In S. Granger, J. Hung, & S. Petch-Tyson (Eds.), *Computer learner corpora, second language acquisition and foreign language teaching* (pp. 143–73). Amsterdam, Netherlands: John Benjamins.

Hellermann, J. (2009). Practices for dispreferred responses using *no* by a learner of English. *International Review of Applied Linguistics, 47*, 95–126.

Hunston, S. (2002). *Corpora in applied linguistics*. New York, NY: Cambridge University Press.

Hyland, K. (2004). Disciplinary interactions: Metadiscourse in L2 postgraduate writing. *Journal of Second Language Writing, 13*(2), 133–51.

Maden-Weinberger, U. (2008). Modality as indicator of L2 proficiency? A corpus-based investigation into advanced German interlanguage. In: M. Walter & P. Grommes (Eds.). *Fortgeschrittene Lernvarietäten* (pp. 141–64). Tübingen, Germany: Niemeyer.

Maynard, C., & Leicher, S. (2007). Pragmatic annotation of an academic spoken corpus for pedagogical purposes. In E. Fitzpatrick (Ed.), *Corpus linguistics beyond the word: Corpus research from phrase to discourse* (pp. 107–15). Amsterdam, Netherlands: Rodopi.

Myles, F., Mitchell, R., & Hooper, J. (1999). Interrogative chunks in French L2: A basis for creative construction? *Studies in Second Language Acquisition, 21*(1), 49–80.

Nesselhauf, N. (2004). How learner corpus analysis can contribute to language teaching: A study of support verb constructions. In G. Aston, S. Bernardini, & D. Stewart (Eds.), *Corpora and language learners* (pp. 109–24). Amsterdam, Netherlands: John Benjamins.

Pavlenko, A., & Driagina, V. (2007). Russian emotion vocabulary in American learners' narratives. *Modern Language Journal, 91*(2), 213–34.

Ramirez-Verdugo, D., & Romero-Trillo, J. (2005). The pragmatic function of intonation in L2 discourse: English tag questions used by Spanish speakers. *Intercultural Pragmatics, 2*(2), 151–68.

Reinhardt, J. (2007). *Directive usage by ITAs: An applied learner corpus analysis* (Unpublished doctoral dissertation). Pennsylvania State University, University Park.

Rundell, M. (Ed.). (2007). *Macmillan English dictionary for advanced learners* (2nd ed.). Oxford, England: Macmillan Education.

Skiba, R., Dittmar, N., & Bressem, J. (2008). Planning, collecting, exploring, and archiving longitudinal L2 data. In L. Ortega & H. Byrnes (Eds.), *The longitudinal study of advanced L2 capacities* (pp. 73–88). Mahwah, NJ: Erlbaum.

Suggested Readings

Aijmer, K. (2009). *Corpora and language teaching*. Amsterdam, Netherlands: John Benjamins.

Aston, G., Bernardini, S., & Stewart, D. (2004). *Corpora and language learners*. Amsterdam, Netherlands: John Benjamins.

Belz, J. A. (2007). The role of computer mediation in the instruction and development of L2 pragmatic competence. *Annual Review of Applied Linguistics, 27*, 45–75.

Fitzpatrick, E. (Ed.). *Corpus linguistics beyond the word: Corpus research from phrase to discourse*. Amsterdam, Netherlands: Rodopi.

Ortega, L., & Byrnes, H. (Eds.). (2008). *The longitudinal study of advanced L2 capacities*. Mahwah, NJ: Erlbaum.

Romero-Trillo, J. (2008). *Pragmatics and corpus linguistics: A mutualistic entente*. Berlin: De Gruyter.

Simpson, R. C., Briggs, S. L., Ovens, J., & Swales, J. M. (2002). *The Michigan Corpus of Academic Spoken English*. Ann Arbor: The Regents of the University of Michigan.