Let's Get SIRIous! Voice Recognition in Language Learning

James Hunter
Gonzaga University

Voice recognition, or automatic speech recognition (ASR), technology is now widely available at little or no cost, and it shows promise in language education, primarily in the area of pronunciation training, where research suggests the technology can outperform human teachers (Golonka, Bowles, Frank, Richardson, & Freynik, 2014, p. 88). This paper discusses how ASR works, how accurate it is, and how it can be applied to language learning both in and out of the classroom. This discussion is followed by a brief examination of recent research into the pedagogical applications of the technology, and concludes that there are many valid reasons to incorporate ASR technologies into language instruction.

James Hunter (hunter@gonzaga.edu) has taught ESL/EFL for over 20 years and is the director of TESOL Programs at Gonzaga University. He has a PhD in Applied Linguistics from the University of Birmingham, UK, and his research interests include second language acquisition, corpus linguistics, instructional technologies, and teacher development.

Voice recognition, or automatic speech recognition (ASR), software has been commercially available for 25 years, but its use in language teaching has been limited. However, the advent of smartphones with bundled voice recognition technology, such as Apple's Siri or Samsung's Voice, has brought the technology within reach of millions of individuals. As language teachers, we often look for low-cost ways to enhance our instruction and engage learners, and ASR technology is a promising development. This paper will look at ASR from a language teaching and learning perspective and will suggest useful ways in which it can be incorporated in and outside the classroom to augment instruction.

A few preliminary points are worth noting, however. First, all ASR technologies that are currently available for free (i.e. not commercial software such as Dragon NaturallySpeaking or Rosetta Stone) require an Internet connection. The reasons for this will be explained below, but it is important to note that the examples given in this paper all require a fast Internet connection. Second, while many of the examples below feature Apple's Siri and Google Dictation, these products are not specifically endorsed here; ASR technology is developing very rapidly, and new and better options are becoming available every year.

What is ASR?

The evolution of ASR parallels new developments in linguistics in many ways: ASR relies heavily on probabilistic (statistical) and usage-based assumptions about language production and comprehension, which are also theoretical cornerstones of corpus linguistics (see, for example, Leech, 1991). In this regard, linguistics (in the US, at least) has moved from a strict Chomskyan approach, which has always remained firmly in opposition to usage-based approaches to linguistic theory. As Chomsky (1969) put it, "It must be recognized that the notion of 'probability of a sentence' is an entirely useless one, under any known interpretation of this term" (p. 57). Applied linguistics, on the other hand, and computational linguistics in particular, have largely followed a different trajectory and have made considerable contributions not only to linguistic theory but more importantly, to daily life. Two examples of how probabilistic models have achieved evident success are search engines and machine translation. Granted, the latter still leaves much to be desired—polysemy, especially in (idiomatic) lexical bundles, renders many automatic translations quite inadequate for practical purposes, offering little more than gist—but it is hard to argue with the fact that people are getting what they need from search engines. Google alone had over 1.2 trillion hits in 2012 ("Google Search Statistics," n.d.), and according to Mitchell (2012), 16%–20% of Google searches on any given day have never been asked before. The key to handling novel search strings lies in employing the likelihood of a given word or string of words (a probabilistic approach) together with algorithms to determine the relevance of the results.

ASR employs a number of processes, some algorithmic but most probabilistic, to recognize continuous speech. In brief (see Jurafsky & Martin, 2009, pp. 232–82 for a comprehensive explanation), the stages in the recognition process are:

1. The sound of one's speech is recorded by the device's microphone, cleaned up to eliminate extraneous background noise, and digitized (converted to a digital file), which is then compressed and sent to cloud servers for processing.
2. The sound wave is scanned for phonemes; when a phoneme, such as the /d/ of *do* is recognized, the statistical likelihood of subsequent phonemes or silence is used to

calculate the accuracy of that /d/ actually being a /d/. The point here is that the process is recursive, not linear; the system has to backtrack continually to ensure comprehension—just as people do when they have to decipher speech under noisy conditions, when encountering unfamiliar accents, or when dealing with homophones; for example, disambiguating “super salad” from “soup or salad,” or “Gladly, the cross I’d bear” from “Gladly, the cross-eyed bear.”

3. As the string of probable phonemes is recognized, segmentation into legitimate—and probable—words becomes increasingly possible. It is important to bear in mind that any given utterance is not composed of individual words with silences between them; in fact, much of the silence in an utterance comes from stops *within* words, rather than between them. To demonstrate this, Figure 1 shows part of the previous sentence displayed as a waveform using the Praat software (Boersma and Weenink, 2011). As can be seen, silence (the absence of vertical waveform in the top third of Figure 1) does not correspond to word boundaries.

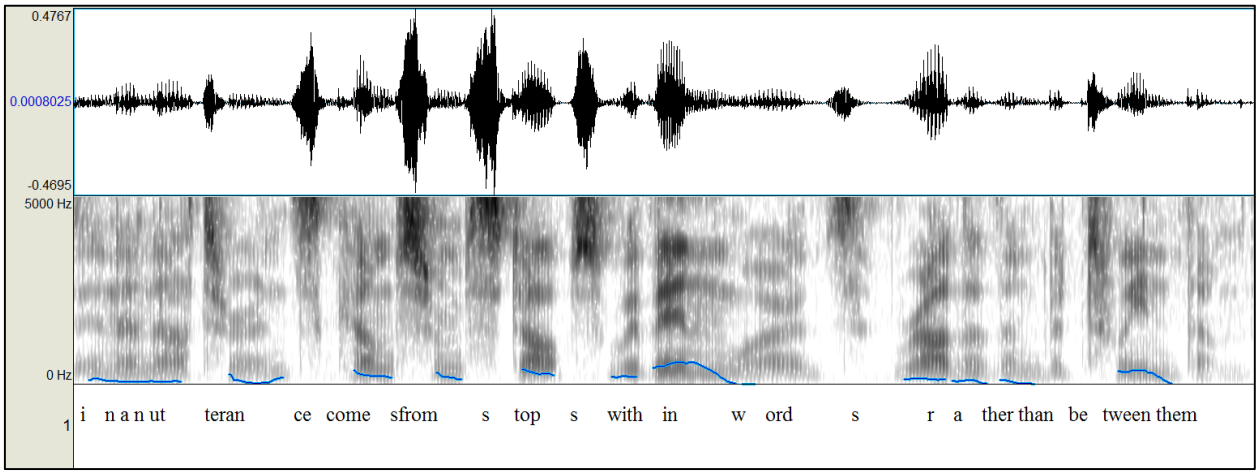

Figure 1: Praat analysis of a phrase showing silence (stops) between words.

4. As words and strings of words are assembled and transcribed, they are verified again for likelihood of co-occurrence, this time using statistical information from vast collections of text (corpora). This is one reason Siri requires an Internet connection: probabilistic inferences require either training (in the sense of teaching the software to recognize a single user’s voice) or access to very large acoustical databases in order to match input to likely phonemes, words, and phrases—and that is simply too much data to store on a smartphone.

5. The transcription is sent back to the device of origin and appears in real time (i.e., it shows up as soon as it is recognized, or, in the case of question answering apps like Siri, the answer is provided).

What is truly remarkable about the above process is the speed at which it occurs: ASR applications can now comfortably deal with dictation speeds of 100+ words per minute (wpm), in contrast to typical typing speeds, which average between 50 and 80 wpm (Harwath, Gruenstein, & McGraw, 2014) The speed and accuracy with which today’s ASR technologies can operate are what make them interesting pedagogically.

How Accurate is ASR?

To test the accuracy of iPhone ASR and Google's dictation.io website, I first made an mp3 recording of a short text (see Appendix) read at slightly slower than normal speed, about 100 wpm. Using the Notes application, or app, I then played it through inexpensive PC speakers into the iPhone (once "Enable Dictation" is turned on in the phone's settings, a microphone icon will appear next to the spacebar on the on-screen keyboard). I had to pause the recording three times to allow the ASR app to catch up, but even under these conditions, the dictation software achieved 97% accuracy, measured in terms of the number of words either substituted or omitted, divided by the total number of words. A colleague read the same text directly into the iPhone, and the ASR achieved 98% accuracy. With accuracy rates this high—Google currently claims an accuracy rate of 92%, while Apple boasts a 95% accuracy rate (Novet, 2015)—it would seem reasonable to assume that a learner could use the technology to identify problematic areas in her pronunciation and be reasonably confident that any substitutions or omissions could be attributed to her output, not the ASR software. To test this assumption, I asked the 24 students in my intermediate Oral Communication course to read the same text into their phones and send me the results, which were then manually compared to the original (see Figure 2), with omissions, additions, and substitutions highlighted in red. In this example, the student (Arabic L1) achieved 70% accuracy.

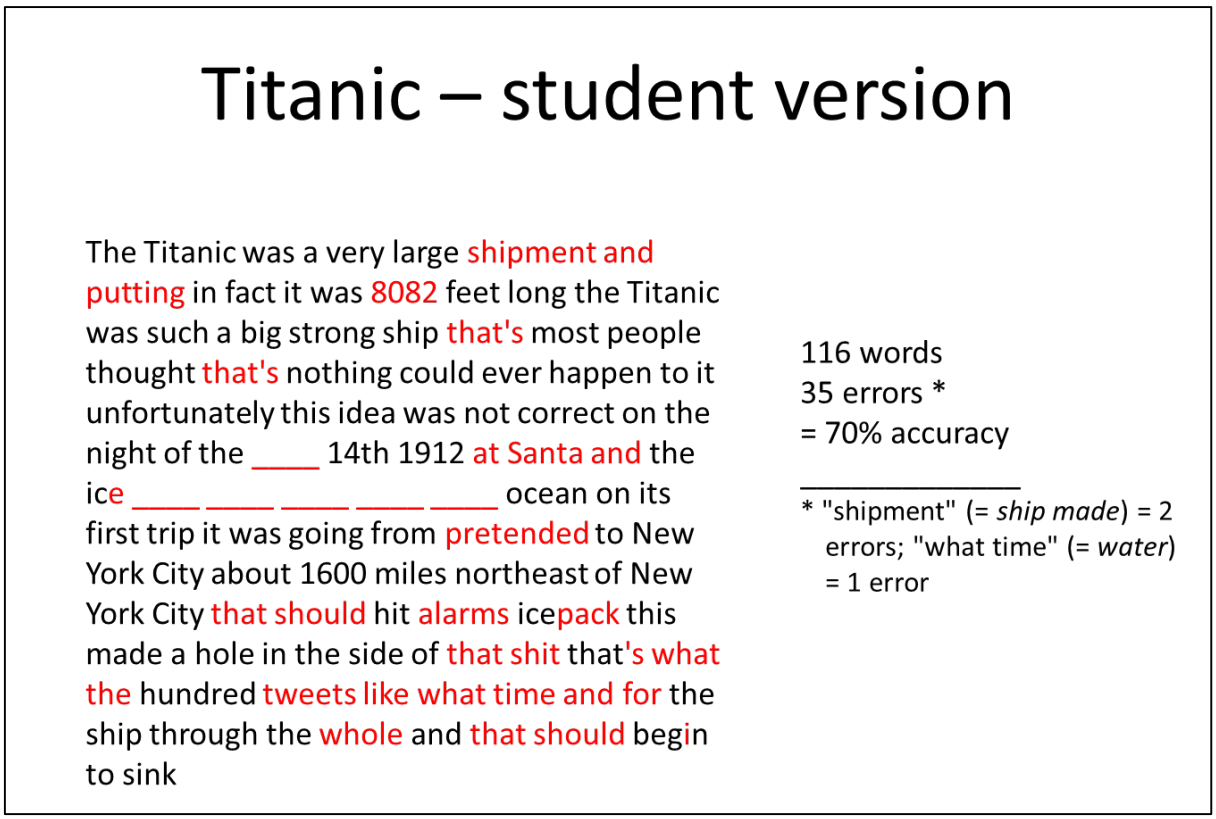

Figure 2: ASR transcription of a student recording of *Titanic* text, showing errors in red.

It should be noted that this accuracy calculation was post hoc: the students only saw the words appearing on their screens. One might think that this low recognition accuracy rate would be disheartening, but what I found when introducing the idea of practicing pronunciation using ASR technologies to my students was that they spent far more time huddled over their phones, both in and out of class, trying to get it right than they would ever have committed to traditional pronunciation drills and exercises.

Interestingly enough, one of the misrecognitions that occurred in all three versions was the following:

original: *Water entered the ship through the hole, and the ship began to sink.*

ASR version: *Water entered the ship through the* ***whole****, and the ship began to sink.*

The software has to pick a transcription for the homonyms *hole/whole*, and despite the fact that the word *hole* appears in the text shortly before this point, it chose *whole*. This exemplifies one crucial feature of current ASR: it does not pay attention to context beyond the sentence level. It selects the most statistically relevant interpretation of a word or phrase based on frequency in everyday discourse, not the interpretation that is most coherent in that particular text. A search in the Corpus of Contemporary American English (Davies, 2009) for the phrases *through the hole* and *through the whole* confirms that the latter is about twice as frequent, while a Google search shows it to be almost 13 times more common.

To borrow a term from one theory of communicative competence, it could be said that ASR lacks discourse competence (Hymes, 1972). An interesting question is whether humans have anything akin to statistical knowledge about frequency of occurrence of phonemes, words, collocations, and multiword sequences (see Granger & Paquot, 2008, p. 42). Many contemporary linguists would argue that we indeed do. Taylor's (2012) *The Mental Corpus* makes a compelling argument for this case, as does Patricia Kuhl's (2010) TED Talk, *The Linguistic Genius of Babies*, which argues that from the very earliest age, babies are "taking statistics" on the occurrence of language-specific phonemes in the speech around them. This usage-based theory of linguistic competence is without doubt more productive than one which separates syntax and semantics, and in fact points out the weakness in Chomsky's argument against probabilistic accounts and corpus linguistics in general: he insists on the *sentence* as the relevant unit of analysis, whereas ordinary speech and writing becomes far more predictable once the *idiom principle*, proposed by Sinclair (1991), is taken into account:

> A language user has available to him or her a large number of semi-preconstructed phrases that constitute single choices, even though they might appear to be analyzable into segments. To some extent this may reflect the recurrence of similar situations in human affairs; it may illustrate a natural tendency to economy of effort; or it may be motivated in part by the exigencies of real-time conversation. (p. 110)

The pedagogical implications of this perspective are far-reaching, but at a minimum argue for teaching practices and materials which emphasize single choices (formulaic speech, collocations, phrasemes) rather than the traditional words-and-rules approach found in the majority of classrooms and textbooks.

ASR in and out of the Classroom

Speaking

Pronunciation training and accent reduction are probably the most obvious applications of ASR technology in language learning, and now that the technology can transcribe natural speech with levels of accuracy which are quite acceptable, there are good reasons to incorporate it into regular language learning activities. There is a major motivational advantage to the use of ASR as a pronunciation feedback system: the feedback is immediate, since the (mis)understood words and phrases are transcribed in real time (Ahn & Lee, 2015). In contrast, a human listener will generally deal with misunderstandings by continuing to listen, in the hope that further context will provide clues as to intended meaning. This is desirable in terms of developing fluency and confidence, but in terms of focusing learner attention at the level of word stress, linking, and segmental, it is not of much help. Teachers often tell their students that phonemic distinction is important because meaning can be affected, but the reality is that a student who says "dare" instead of "there" or "tree" instead of "three" is unlikely to encounter listeners who simply refuse to interpret the words with latitude. ASR, as we have seen, will do just that. A Korean student of mine was recently trying to say the phrase "though there are . . ." and could only get the technology (dictation.io) to transcribe "do terra." This was a source of frustration, for sure, but the student sought me out during my office hours for a targeted pronunciation lesson, just to try to beat the machine. Similarly, my Arabic L1 students who say "paper" in a way that sounds to me like "baber" will actually see "Baber" transcribed (presumably because the software concludes it must be a name) and "botatoes" transcribed not as "potatoes" but as "but it does."

If, as I am suggesting, ASR technology can motivate students to notice, focus on, and practice phonological challenges, it also does so tirelessly and with infinite patience, which is more than I can claim for myself. This means that learners can "communicate," in the sense of getting a message across, and get feedback on the success of the communication without any of the affective epiphenomena that accompany human interaction, and above all, without judgment or embarrassment, and at their own pace. In fact, the technology seems to replicate a willingness to understand in spite of pronunciation issues, in other words, a forgiving ear. It can do this precisely because it refers to probability of occurrence of certain language strings in the real world. So, for instance, if you were to say the sequence "boot eat does," the resulting transcription is more likely to be "but it does" because this is a far more likely sequence in the real world, by a factor of 300 million (a quick search on Google gives 0 hits for the former and 317 million for the latter).

Virtual assistant applications like Siri, Cortana, and Google Now can be used in many other ways to practice speaking. I give teams of students a worksheet of facts to find out about the world (see Figure 3) by asking Siri questions. In this instance, Siri reads the answer aloud, assuming it understands the question, and the students have to be able to give the answer verbatim (they can get the answer as many times as they like by saying "repeat").

Figure 3: Worksheet of information to research using Siri or Google Voice Search.

Writing

Beginner-level students, and especially those who are challenged by English spelling, can benefit from a simple exercise in which they read a word on a flash card and see if the ASR transcribes the same word. The point here is that they will not see the non-words they might produce on a spelling test. More advanced students, in contrast, can dictate rather than write or type first drafts of essays or other writing assignments, just to get their thoughts down in writing. The advantage of this approach is that it allows for more fluency—as the writer does not have to stop to worry about the spelling of a challenging word, as long as she can say it reasonably comprehensibly—and also provides an inherent reason for editing work (to verify that the ASR transcribed what was intended). With the addition of text-to-speech technology, which is built into most smartphones and computers, writers can hear their words read back to them in a reasonable facsimile of a human voice, which offers another opportunity to compare the intended meaning with the transcribed text, as well as further reinforcing the pairing of written word with sound.

Both of these suggestions may appear to promote less, not greater, attention to spelling conventions among English learners. But it also seems to be the case that American college students, at least, are over-reliant on spelling technologies and tend to assume that they eliminate the need for careful proofreading (Galletta, Durcikova, Everard, & Jones, 2005), so English language learners are in good—or bad—company. On the other hand, *anything* which motivates language learners to talk more, write more, and read more is worth investigating, and I look forward to seeing studies that investigate the relationship between ASR use and literacy skills.

Other Applications of ASR in Language Learning

Possibly the most exciting developments in instructional technology at present involve the combination of ASR and artificial intelligence technologies to create virtual learning environments and characters with which learners can interact. Macedonia, Groher, and Roithmayr (2014), for example, claim that “intelligent virtual agents” such as their “Billie,” a virtual character which teaches learners vocabulary with the help of “iconic gestures,”

outperform human trainers (p. 1). Similarly positive results are claimed for virtual scenarios, in which learners interact with virtual characters to complete communicative goals such as purchasing rail tickets (Chiu, Liou, & Yeh, 2007; Morton, Gunson, & Jack, 2012). Coniam (1998) suggested that ASR scoring of texts read by learners could represent a useful automated assessment tool, although he concluded at the time that a "robust generic speech model is not yet available" (p. 20), meaning that the ASR needed to be trained to recognized individual speech, which is no longer the case. At my university, we are developing a web-based application of our "Small Talk" database (Hunter, 2011) which is part of a system of delayed corrective feedback of learners' oral production. The web application will incorporate ASR so that learners will be able to check the accuracy of their reformulations, in particular their pronunciation, against those provided by their teacher. These are just a few of the ways in which researchers and practitioners are harnessing the technology to facilitate or improve upon language pedagogy.

Conclusion

In their review of 350 studies of technology in language teaching, Golonka, Bowles, Frank, Richardson, and Freynik (2014) conclude that the most impressive contributions of ASR to language instruction are in the area of pronunciation training:

> Among those technologies that were included in our review, the only strong support we found for an impact of technology on FL [foreign language] learning and teaching were for the ASR programs and chat. Research shows that the ASR technology can facilitate improvement in pronunciation to a larger extent than human teachers can and, because of constant improvements of this technology, ASR programs have great potential in FL learning. (p. 88)

For teachers who struggle to find the time and means to give their students the individualized pronunciation training they need, this is good news, and I hope that I have made the case that there are good reasons to incorporate ASR-based activities into our regular teaching, with a view to encouraging learners to build these tools into their language learning strategies and practices. It is time for teachers to get serious about ASR.

References

Anh, T. Y. & Lee, S. M. (2015). User experience of a mobile speaking application with automatic speech recognition for EFL learning. *British Journal of Educational Technology*, *47*(4), 778–786. doi:10.1111/bjet.12354

Boersma, P. & Weenink, D. (2011). Praat [Computer software]. Retrieved from: http://www.fon.hum.uva.nl/praat/

Chiu, T., Liou, H., & Yeh, Y. (2007). A study of web-based oral activities enhanced by automatic speech recognition for EFL college learning. *Computer Assisted Language Learning, 20*(3), 209–233.

Chomsky, N. (1969). Some empirical assumptions in modern philosophy of language. In S. Morgenbesser, P. Suppes, & M. White (Eds.), *Philosophy, science and method: Essays in honor of Ernest Nagel*. New York, NY: St. Martin's Press.

Coniam, D. (1998). The use of speech recognition software as an English language oral assessment instrument: An exploratory study. *CALICO Journal*, *15*(4), 7–23.

Davies, M. (2009). The 385+ million word corpus of contemporary American English (1990–2008+): Design, architecture, and linguistic insights. *International Journal of Corpus Linguistics*, *14*(2), 159–190.

Galletta, D. F., Durcikova, A., Everard, A., & Jones, B. M. (2005). Does spell-checking software need a warning label? *Communications of the ACM*, *48*(7), 82–86.

Granger, S., & Paquot, M. (2008). Disentangling the phraseological web. In S. Granger & F. Meunier (Eds.), *Phraseology: An interdisciplinary perspective* (pp. 27–50). Amsterdam: John Benjamins Publishing.

Golonka, E. M., Bowles, A. R., Frank, V. M., Richardson, D. L., & Freynik, S. (2014). Technologies for foreign language learning: A review of technology types and their effectiveness. *Computer Assisted Language Learning*, *27*(1), 70–105.

Google search statistics. (n.d.). Retrieved March 10, 2016 from Internet Live Stats: http://www.internetlivestats.com/google-search-statistics/

Harwath, D., Gruenstein, A., & McGraw, I. (2014). Choosing useful word alternates for automatic speech recognition correction interfaces. *INTERSPEECH* – 2014 (949–953).

Hunter, J. (2011). 'Small Talk': Developing fluency, accuracy, and complexity in speaking. *ELT Journal*, *66*(1), 30–41.

Hymes, D. (1972). On communicative competence. In J. B. Pride & J. Holmes (Eds.), *Sociolinguistics: Selected readings* (pp. 269–293). Harmondsworth, UK: Penguin.

Jurafsky, D., & Martin, J. H. (2009). *Speech and language processing: An introduction to natural language processing, computational linguistics, and speech recognition.* Englewood Cliffs, NJ: Prentice Hall.

Kuhl, P. (2010, October). *Patricia Kuhl: The linguistic genius of babies* [Video file]. Retrieved from https://www.ted.com/talks/patricia_kuhl_the_linguistic_genius_of_babies ?language=en

Leech, G. (1991). The state of the art in corpus linguistics. In K. Aijmer & B. Altenberg (Eds.) *English corpus linguistics: Studies in honour of Jan Svartvik* (pp. 8–29). London, UK: Longman.

Macedonia, M., Groher, I., & Roithmayr, F. (2014). Intelligent virtual agents as language trainers facilitate multilingualism. *Frontiers in psychology*, *5*, (1–4).

Mitchell, J. (2012, February 29). How Google search really works. *Readwrite*. Retrieved from http://readwrite.com/2012/02/29/interview_changing_engines_mid-flight_qa_with_goog/

Morton, H., Gunson, N., & Jack, M. (2012). Interactive language learning through speech-enabled virtual scenarios. *Advances in Human-Computer Interaction*, *2012*, 1–14.

Novet, J. (2015, June 8). Apple claims Siri's speech recognition tech is more accurate than Google's. Retrieved from http://venturebeat.com/2015/06/08/apple-claims-siris-speech-recognition-tech-is-more-accurate-than-googles/

Sinclair, J. (1991). *Corpus, concordance, collocation*. Oxford, UK: Oxford University Press.

Taylor, J. R. (2012). *The mental corpus: How language is represented in the mind*. Oxford, UK: Oxford University Press.

Appendix

Titanic text (original)

The Titanic was a very large ship made in Britain. In fact, it was 882 feet long. The Titanic was such a big, strong ship that most people thought that nothing could ever happen to it. Unfortunately, this idea was not correct. On the night of April 14th, 1912, it sank in the icy water of the North Atlantic Ocean on its first trip. It was going from Britain to New York City. About 1,600 miles northeast of New York City, the ship hit a large iceberg. This made a hole in the side of the ship that was 300 feet long. Water entered the ship through the hole, and the ship began to sink.

