- MENU
- PRINTABLE VERSION
- HELP & FAQS

Language as Chunks, not Words

Ramesh Krishnamurthy

Cobuild, University of Birmingham

Many people think of language as words. Words are small, convenient units, especially in written English, where they are separated by spaces. Dictionaries seem to reinforce this idea, because entries are arranged as a list of alphabetically-ordered words. Traditionally, linguists and teachers focused on grammar and treated words as self-contained units of meaning, which fill the available grammatical slots in a sentence. More recently, attention has shifted from grammar to lexis, and from words to chunks. Dictionary headwords are convenient points of access for the user, but modern dictionary entries usually deal with chunks, because meanings often do not arise from individual words, but from the chunks in which the words occur. Corpus research confirms that native speakers of a language actually work with larger "chunks" of language. This paper will show that teachers and learners will benefit from treating language as chunks rather than words.

1. Written language as "chunks"

When children learn English as L1, they are first taught to recognize the letters of the alphabet, before learning to read and write words, e.g. *n — — — o — — — t = "not", f — — — o — — — r = "for", t — — — h — — — e = "the"*, and so on. But once they have progressed to the word level, they don't continue to read or write texts letter by letter, because this would make the processing both very slow and very difficult. Letters are basic units of language, but they are only operational units for language processing at the very initial stage of language learning.

For many people, words are the most obvious unit of language. In written English, this is emphasized by the fact that words are separated by spaces. Kurtus (2001) claims: "Most people read one word at a time, saying the word to themselves." But this cannot be true.

If we really processed language one word at a time, communication would still be very slow and very difficult, e.g. *Not — — — for — — — the — — — first — — — time — — — , — — — an — — — argument — — — had — — — broken — — — out — — — .*

If we read this text one word at a time, we first need to process WORD ONE, recall and keep in mind all of its possible uses or meanings, then process WORD TWO in a similar way, then go

back to WORD ONE and see if we can now decide which use or meaning of WORD ONE was intended and how WORD TWO relates to WORD ONE; if this is still not clear, we would need to process WORD THREE, and so on, keeping an enormous number of unresolved possibilities in our minds for a long time.

It is very unlikely that words are the operational unit for native-speakers (Fostering Second Language Development in Young Children 1995, Kendon 1996, Ellis 1997), and therefore should not be for advanced learners of the language (Porto 1998, Ketko 2000, Markus 2000, TOEFL Strategy #1 2002).

Words are in fact just another intermediate unit of language, like letters. The real operational unit of most native-speakers is "chunks", groups of words that form meaningful units, e.g. *Not for the first time, — — — an argument had broken out — — — over breakfast — — — at number four, Privet Drive — — — . — — — Mr Vernon Dursley had been woken — — — in the early hours of the morning — — — by a loud, hooting noise — — — from his nephew Harry's room — — —* . (Rowling 1998).

Therefore we should be helping advanced learners of the language to move towards operating at the "chunk" level as well. "Chunk-by-chunk" processing makes communication faster, more efficient, and easier for mutual comprehension. The text in the example contains 217 characters, 40 words, but only 8 chunks.

2. Spoken language as "chunks"

Very similar developments take place in spoken language. One source (*Fostering Second Language Development in Young Children,* 1995) says: "While children may appear to be making more mistakes during experimentation, they are actually learning to internalize chunks of appropriate speech. They test these chunks of language by using them in situations that may or may not be appropriate. The feedback they receive helps them determine whether they have guessed correctly." After the initial stage of learning, we don't continue to process spoken language phoneme by phoneme or syllable by syllable. We speak in chunks called "tone units" (also called intonation units, or breath groups), with pauses in between to breathe, to allow for a response, or for emphasis.

The Language Fun Farm (Interview with David Horner, 2002) focuses on listening, and segmenting discourse into chunks, rather than constructing chunks during production (see Section 7). However, Ketko (2000) addresses the crucial role of multi-word chunks in facilitating communicative competence.

3. Why call them "chunks" rather than "phrases"?

The term "phrase" has a long history of technical usage in linguistics, and means different things to different people, depending on which linguistic theory they are working with. "Chunks" is a less established term, but therefore has less history, less "baggage" associated with it.

4. Why "chunks" rather than longer units, such as "clauses" or "sentences"?

The exact dimensions and attributes of "chunks" as language processing units have not yet been firmly established. However, I would argue firstly that a chunk is primarily a lexical unit, and may represent units at various functional and formal levels in the grammar hierarchy; secondly, that it is a unit of memory; and thirdly, that it is necessarily variable in length, but is unlikely to be longer than a clause-element, especially for written texts, where clauses and sentences may be very long.

Both "clause" and "sentence" are grammatical units, and therefore require grammatical processing and comprehension at a higher or more abstract level, which may or may not be carried out subsequently, after the initial lexical processing. In my intuitive and instinctive division of a written text into chunks (or meaningful units), the chunks can be analysed in various ways at different grammatical levels, e.g. the first chunk *Not for the first time,* can be analysed functionally as an adjunct of frequency, or formally as a negative particle with a prepositional phrase consisting of a preposition and a noun phrase (consisting of the definite article + adjective + noun); and the chunks vary in length from 2 words to 6. The chunks are grammatical units, but vary in level from clause-element to sentence.

However, it is not the grammatical level of the unit that is important in this initial stage of language processing, but the length of the unit. Psychologists have suggested a maximum length for information processing: George Miller suggested "that there is a limit to how much information that a person can remember. This is immediate memory, single dimensional information such a series of numbers, tones, or events. This limit is seven, give or take two." (Crow et al 1998). Applin (1999) claims that "This proposition has been verified at all levels of cognitive processing by independent research."

Corpus research suggests that collocation also operates within a span of about 5 words: Sinclair et al (1970: 9) said: "Collocation, or significant co-occurrence of lexical units, assumes that the extent of the environment, the "co-", can be specified…Later investigation …showed that the optimum extent (called span) was four words on either side of the node. A shorter span would miss valuable evidence, and a longer one would overlay the relevant patterns with more distant material."

Kurtus (2001) notes: "A newspaper column usually has 4 or 5 words per line". In fact, my cursory examination of several newspapers indicates 5-10 words per line, and of course the lines are divided equally rather than into meaningful units, but this may still be indicative that native-speakers can process 5-10 words at a time. Another web source (Plain English at Work, 1997) says: "There's a limit to the number of words that readers can comfortably follow in a line of type. If the lines are too long, readers tend to lose track. But if the lines are too short, the reading flow is interrupted too often." And the Plain English Campaign (The plain English guide to design and layout, 2003) echoes this: "Line length can affect the ease and speed of your reading. Very long and very short lines force you to read more slowly. The size of the type you should use depends on the length of the line. Longer lines of body text need larger type. It is helpful to think of line length in terms of the number of characters in the line (including spaces). A line of body text should normally contain 60 to 72 characters, or about 10 to 12 words."

Of course, native-speakers may in fact sometimes work with longer and higher-level units of language, both in writing and in speech. Many school and college websites recommend "rapid reading", "skim reading", or "scanning" to their students, e.g. Kentwell (2002). But this is a different process—note that not every word or sentence is being read.

For spoken language, Kendon (1996) says: "in a continuous discourse, speakers group tone units into higher order groupings and so we can speak of a hierarchy of such units…a series of tone units linked intonationally or by an absence of pauses into a coherent higher order grouping…tone units are organized e.g. by intonation patterns, types of pauses, by subordination to one another, etc."

5. Dictionaries and chunks

Dictionaries look like lists of words, with information attached to each word, e.g. *merger, meridian, meringue, merit*, so this unfortunately reinforces people's belief that the basic unit of language is words. But in fact, dictionary entries are arranged by headwords only for convenience of access, to take advantage of alphabetical order. And even at the headword level (especially in EFL dictionaries) we are presented with chunks, not individual words, much of the time, e.g. *first name*, *first night, first offender, first-past-the-post*.

Within entries, EFL dictionaries focus even more on chunks rather than words, as we will see if we look up some of the words in the (Rowling 1998) text used earlier (I have underlined the chunks). Collins COBUILD English Dictionary for Advanced Learners (CCEDAL, 2001) usually shows the chunk in its definitions as well as in examples, e.g. **first 3:** When something happens or is done for the **first** time, it has never happened or been done before. **time 13:** *House prices are rising for the first time since November.* **[3] over 5:** If something happens **over** a particular period of time or **over** something such as a meal, it happens during that time or during the meal. *...Over breakfast we discussed plans for the day.* EFL dictionaries often highlight the chunks in bold type, e.g. the Oxford Advanced Learner's Dictionary (OALD, 2000): **hour 8: the small/early hours** (also **the wee small hours** *ScotE, AmE* also **the wee hours**) ... *The fighting began in the early hours of Saturday morning*. Different dictionaries give different amounts of information, and often present similar information in different ways, e.g. the Longman Dictionary of Contemporary English (LDOCE, 2001): **over[1] 10** during: *Will you be home over the Christmas vacation?| Over a period of ten years he stole a million pounds from the company.* | *Can we talk about this over dinner?*

Unfortunately, none of the above dictionaries deals specifically with the chunk *an argument had broken out*; at the entry for *argument*, CCEDAL has *set out arguments, convince by argument, trigger arguments, cause heated argument, get into an argument;* LDOCE and OALD show *have an argument, get into an argument, win/lose an argument*. And at the entry for *break out*, CCEDAL gives *war, fighting, disease, fight;* LDOCE gives *war, fire, disease, something unpleasant, scuffles;* and OALD gives *war, fighting, unpleasant events, fire*.

Sometimes, one chunk depends on another for its meaning. For example, if you look up *hoot* in the text: *Mr Vernon Dursley had been woken in the early hours of the morning by a loud, hooting noise from his nephew Harry's room*, you will not know which of the uses/meanings given in the dictionary is intended, e.g. CCEDAL **hoot 1** If you **hoot** the horn on a vehicle or if it hoots, it makes a loud noise on one note. **2** If you **hoot**, you make a loud high-pitched noise when you are laughing or showing disapproval. **3** When an owl **hoots**, it makes a sound like a long `oo'. You have to read a few more lines of the text: *'Third time this week!' he roared across the table. 'If you can't control that owl, it'll have to go!'* before you realize that use/meaning 3 relating to owl noises is the one you need.

6. Corpus evidence for chunks

Several current EFL dictionaries are based on a corpus (a large collection of authentic language texts stored in a computer and investigated by sophisticated software), and select the chunks to be included on the basis of their frequency in the corpus. For example, in the Bank of English corpus (450 million words of text) at the University of Birmingham, there are over 1.5 million examples of *not*, nearly 4 million examples of *for*, nearly 25

million examples of *the*, 621,000 examples of *first*, and 706,000 examples of *time*. There are 45,000 examples of *the first time*, of which 31,000 are *for the first time*, and 608 *not for the first time*. Therefore, it is not surprising that the dictionaries include the chunks *the first time* and *for the first time*, but omit *not for the first time*.

On the other hand, out of 24,000 examples for *argument*, only 321 contain *break out*; and out of 7232 examples for *break out*, only 49 contain *argument*, and therefore it is reasonable that none of the dictionaries shows the chunk *argument + break out*.

Corpus evidence also shows us how chunks develop. For example, this advertisement from a computer company has just appeared in UK newspapers: "Who gives you the best exclusive entertainment? THE MOTHER OF ALL BROADBAND SERVICES". In the corpus, we find that Saddam Hussein first used this chunk in 1990, warning the U.S. that the fight for Kuwait would be the "Mother of all Battles" (so the company is obviously cashing in on the current Iraqi crisis).

> …today quoted him as saying, `The mother of all battles has begun and the…
>
> …Saddam Hussein had billed as the mother of all battles.
>
> ….676,000 Allied troops in the 1990 mother of all battles…
>
> ….urged Iraqis to prepare for the mother of all battles.

Very soon, the chunk became used in an extended sense, for any kind of disagreement:

> hideously divisive, and could be the mother of all political battles

Then the word *battle* was dropped, and other words for disagreement were used instead:

> The mother of all business feuds ended when…
>
> …warned the Italians to expect the mother of all confrontations when…

And finally, the chunk lost even the association with disagreements, and is now reduced to "mother-of-all" and widely used with a general superlative meaning, "the biggest, most impressive, most extreme, etc":

> Kernaghan is looking forward to the `mother of all parties' in Dublin next…
>
> Black fur, white stripes. The mother of all skunks. I don't know why but…
>
> …toy department and FAO Schwarz, mother of all toy stores, where Barbie has…
>
> The game which was the mother of all video games comes to Game…

7. Chunks in language teaching and learning

There has been much discussion about the role of chunks in language learning recently. For example, The Language Fun Farm (Interview with David Horner, 2002) says: "It's when you start to notice bits, and you say "Oh yeah, I recognise that". When you start to be able to segment what you can hear then you're starting to make progress. But it's very difficult to notice if somebody hasn't drawn your attention to it…You don't have to know it means something, you just have to know, "Oh, yeah - that's *that chunk*"…then the idea is that the chunks you understand will get bigger and also more frequent and eventually there won't be any space between the chunks and you'll understand everything". Many experts recommend the use of chunks in language teaching. Miller (2002) gives examples of chunk-based activities and refers to Ellis (1997): "(1) People chunk at a constant rate: every time they get more experience, they build additional chunks (2) Performance on the task is faster, the more chunks that have been built that are relevant to the task" and to Porto (1998) on why chunks should be taught. Ketko (2000) discusses the importance of multi-word chunks in facilitating communicative competence. Markus (2000) has apparently had great success in teaching Latin through chunks. Elsewhere (TOEFL Strategy #1, 2002) we find: "Students focus on learning "unanalysed chunks" of words and fixed expressions, which they can then use in real situations. Because the word chunks and fixed expressions are formed with the grammar of the language, students learn the grammar of the language as they are learning chunks and fixed expressions, but in a more intuitive way, just as native speakers of a language learn the grammar of their language."

It is therefore crucial that teachers help learners to recognize language as chunks, and that students learn and remember language as chunks. A major advantage is that whereas words can seem to have many meanings (as implied in dictionaries), chunks tend to have only one or two meanings. Teachers must also encourage students to practice producing chunks in their written and spoken output. Of course, words are still useful, but only as access points to chunks.

References

Applin, A. (1999). *The application of language acquisition theory to programming concept instruction: chunks versus programs from scratch.* SIGCSE Doctoral Consortium, New Orleans. Retrieved 11/12/02 from the World Wide Web: <www.cs.utexas.edu/users/csed/doc_consortium/DC99/applin-abstract.html>

CCEDAL (2001). *Collins COBUILD English Dictionary for Advanced Learners* (3rd ed.). Glasgow: HarperCollins Publishers.

Crow, D., Jansen, J., & Orrick, E. (1998). Students: Recommended Readings. *SIGCHI Bulletin, 30 (3).* Retrieved 11/12/02 from the World Wide Web: <www.acm.org/sigchi/bulletin/1998.3/students.html>

Ellis, N.C. (1997). Vocabulary acquisition: word structure, collocation, word-class, and meaning. In Schmitt N. & McCarthy M., *Vocabulary: Description, Acquisition and Pedagogy*. Cambridge: Cambridge University Press.

Fostering Second Language Development in Young Children (1995). Published by ERIC Clearinghouse on Languages and Linguistics. Retrieved 11/12/02 from the World Wide Web: <www.cal.org/ericcll/digest/ncrcds04.html>

Interview with David Horner (2002). Published by The Language Fun Farm (1997-2002). Retrieved 11/12/02 from the World Wide Web: <www.teflfarm.com/teachers/interviews/ 0/ horner/part2.htm>

Kendon, A. (1996). An Agenda for Gesture Studies. *Semiotic Review of Books*, *7(3), 8-12*. Retrieved 11/12/02 from the World Wide Web: <www.univie.ac.at/Wissenschaftstheorie/ srb/srb/gesture.html>

Kentwell, R. (2002). *An Interactive guide to the Research Process*. Published by the Melbourne High School Library. Retrieved 11/12/02 from the World Wide Web: www.mhs.vic.edu.au/home/library/infoproc/skim1.htm

Ketko, H. (2000). Importance of "MultiWord Chunks" in Facilitating Communicative Competence and its Pedagogic Implications. *Language Teacher Online 24.12*. Retrieved 11/ 12/02 from the World Wide Web: <langue.hyper.chubu.ac.jp/ jalt/ pub/tlt/00/dec/ketko.html>

Kurtus, R. (2001). *Reading Faster*. Published by the School for Champions. Retrieved 11/12/02 from the World Wide Web: <http://www.school-for-champions.com/grades/reading.htm>

LDOCE (2001). *Longman Dictionary of Contemporary English* (3rd ed.). Harlow: Pearson Education.

Markus D.D. (2000). *Patterns of Cohesion & Discontinuity as a Teaching Tool for Reading Caesar and Cicero in the Second Year*. Retrieved 11/12/02 from the World Wide Web: <ablemedia.com/pr092700markus.html>

Miller, J (2002). *Presentation Handout* at the Universidad del Valle de Mexico, Mexico City, D.F., Mexico. Retrieved 11/12/ 02 from the World Wide Web: <www.geocities.com/jabbusch/ chunks.htm>

OALD (2000). *The Oxford Advanced Learner's Dictionary* (6th ed.). Oxford: Oxford University Press.

Plain English at Work (1997). Produced by Susan Munter Communications for Australia's Department of Education, Training and Youth Affairs (DETYA) and the Australian National Training Authority (ANTA). Retrieved 09/03/03 from the World Wide Web: <www.detya.gov.au/archive/ publications/plain_en/design.htm>

Porto, M. (1998). Lexical phrases and language teaching. *Forum 36,3.*

Rowling, J.K. (1998). *Harry Potter and the Chamber of Secrets.* London: Bloomsbury.

Sinclair, J., Jones, S., & Daley, R. (1970): English Lexical Studies: Report to OSTI on Project C/LP/08. New Edition: Krishnamurthy (Ed.) 2003. *English Collocation Studies: The OSTI Report*. Birmingham: Birmingham University Press.

The plain English guide to design and layout (2003). Published by the Plain English Campaign. Retrieved 09/09/03 from the World Wide Web: <www.plainenglish.co.uk/ design.html#Anchor-Lin-60872>

TOEFL Strategy #1 (2002). Published by Kaplan Test Prep and Admissions. Retrieved 11/12/02 from the World Wide Web: <www.kaptest.com/repository/templates/ArticleInitDroplet.jhtml?_relPath=/repository/content/International/TOEFL/ Strategy_Sessions/IN_toefl_strat1.html>