ULUSLARARASI

VI. TÜRK DİLİ KURULTAYI BİLDİRİLERİ

20-25 Ekim 2008

Türk Dil Kurumu Yayınları

DERLEM DİL BİLİMİ VE GÜNCEL GELİŞMELER

B. Tahir TAHİROĞLU [*]

1.Giriş

Bilişim ve bilgi teknolojilerinde yaşanan hızlı gelişim, dil bilimi çalışma yöntemlerinde son yıllarda kimi değişimler sağlamıştır. Yapay zeka çalışmaları başta olmak üzere insan dilinin çok önemli bir çözümleme alanı olması ve makinelerin insanla etkileşimi konusunda dille ilgili problemlerin çözümünün giderek artan oranda önem kazanması, dil bilimi ve bilişim bilimlerinin disiplinler arası bir nitelikte buluşmalarını hızlandırmıştır.

Bu bildiride derlem dil biliminin (corpus Linguistics) kısaca tanıtımı amaçlanmıştır. Derlem kavramı ve ilk derlemler ile genel olarak derlemin kullanım alanlarıyla Türkçe için derlem çalışmaları ve son olarak da internette yer alan derlem araçlarından örnekler verilmiştir.

2. Derlem Terimi

Derlem terimi son 20 yılda dil biliminin önemli bir konusu durumuna gelmiştir. Derlem İngilizce literatürde *corpus* olarak geçmekte, Latince beden anlamına gelmektedir. Sözcüğün çoğul biçimi olarak da literatürde *corpora* kullanılmaktadır. Derlem kısaca boyutça büyük metin koleksiyonlarının elektronik biçimlerde bilgisayar ortamında tutulan dilsel kesitler olarak tanımlanabilir. Derlemlerin büyük metin arşivlerinden farkı bu metinlerin sesbilimsel, sözdizimsel ve anlambilimsel olarak bir standart çerçevesinde işaretlenmiş olmasıdır. Derlemler genel ve özel olmak üzere iki ana kategoride ele alınmaktadırlar. Derlemler kendi içinde alana özgü derlemler, referans derlemleri, çok dilli derlemler, paralel, öğrenci derlemleri, art zamanlı ve monitör (çok büyük boyutlu daha çok güncel dili kapsayan derlem) olmak üzere türlere ayrılmaktadır (Baker vd. 2006: 48).

Türkçe literatürde dil bilim araştırmacılarının daha çok corpus terimine *bütünce* karşılığını kullandıkları görülürken bilgisayar bilimi kökenli araştırmacıların *derlem* karşılığını tercih etmektedirler. Bugüne kadar yapılan bilgisayar destekli dil bilimi çalıştaylarında derlem karşılığının alan için terim olarak standartlaşmıştır.

[*] Çukurova Üniversitesi Türkoloji Araştırma ve Uygulama Merkezi

Batı'da derlem çalışmalarının İncil üzerine bağlamlı dizinlerin (concordance) hazırlanmasıyla ortaya çıktığı kabul edilmektedir. Londralı bir kitapçı olan Alexander Cruden'in Concordance'ı İncil için hazırlanan önemli bir kaynak olarak görülmektedir. Bu çalışmada İngilizce kimi sözlerin işlevlerine de değinilmiştir. Bilgisayarların derlem çalışmalarında kullanılmasından önceki dönemlerde, sözlük bilimsel çalışmalar da günümüz derlem dil biliminin ilk yapıtaşlarını oluşturmuştur denilebilir. Bu bağlamda Samuel Jhonson İngilizce sözcüklerin kullanımıyla ilgili bilgileri altı yardımcısıyla Dictionary of the English Language adlı sözlüğün 40 bin maddebaşı sözü için 150 binden fazla tanıklı fiş yazmıştır. Webster'in An American Dictionary of the English Language adlı sözlüğü 1828'de yayımlanmıştır. Bu sözlüğün 1961'de yayımlanan 3. baskısı 10 milyon fişlik bir derlemden oluşmuş yaklaşık 500 bin madde başı sözü içermektedir. Derlem dil biliminin bugün de sözlük biliminde bilgisayar destekli olarak çok önemli bir çalışma disiplini olduğu görülmektedir (Kennedy 1998: 14).

Derlem dil biliminin dil bilimi içinde başlıca bir alan olarak gelişip gelişmeyeceği konusu tartışmalıdır. McEnery ve Wilson (2004), derlem dil biliminin bir ses bilimi, söz dizimi gibi dil bilimi alt dalı olmadığını ancak yöntem bilgisel olarak örneğin söz dizimi ya da biçim bilimi çalışma alanlarına veri sağlayabileceğini belirtmektedirler. Burada şu noktayı da belirtmek gerekir. Bugün derlem çalışmaları hem doğal dil işleme (DDİ) adı verilen daha çok bilgisayar bilimlerinde ele alınan konular için veri sağlamakla birlikte dil bilimi çalışmalarında veriye dayalı sözlüklerin oluşturulmasında sezgisel çıkarımlar yerine kanıta dayalı bir yöntem olarak kullanımını genişletmektedir.

3. Derlem Kullanım Alanları

Büyük boyutlu derlemler özellikle dil öğretiminde başvuru kaynağı olarak düşünülmektedir. Yabancı öğrenciler için dil bilgisi kitaplarının hazırlanmasında, bu kaynaklarda yer alacak kimi eş dizimsel (collocation) örüntülerin gerçek kullanımdaki sıklıklarıyla ortaya konmasında derlemler önemli veriler sunmaktadır. Dil öğrencisi için temel söz varlığına dayalı sözlüklerdeki madde başı sözlerin seçiminde özellikle işaretlenmiş bir derlemin kullanılması gerekmektedir. Yapılandırılmamış, yani işaretleme ve içerik bakımından belgelerin niteliklerinin ortaya konmadığı karmaşık belgelerden sözlük birimlerin bu temel sözlükler için elde edilmesi çok güç olmakla birlikte ortaya çıkan sonucun hatalardan arındırılması da gerekmektedir. Bu bağlamda dil öğretimi sözlüklerinin ve dilbilgisi kitaplarının hazırlanmasında temel olarak tekrarlı birimlerinden temizlenmiş, gürültü olarak kabul edilebilecek söz dışı ögelerin ayrıca ele alınabileceği biçimde hazırlanmış derlemlere gereksinim duyulmaktadır.

Derlemler DDİ çalışmalarında özellikle biçim birimsel çözümlemede istatistiksel yöntemlerin kullanıldığı çalışmalarda bilgisayar için eğitim verisi olarak kullanılmaktadır. Girdi olarak alınan bir yapıdan istenilen doğru biçim birimsel çözümlemede hem kural tabanlı hem de istatistik yöntemde derlem başarım oranını artırmaktadır. Bir makinece okunur sözlüğün (machine readable dictionary) veri tabanı olarak kullanıldığı DDİ uygulamalarında da sözlük birimlerin elde edilmesi

yine derleme dayalı olarak gerçekleştirilmektedir.

Sözlük biliminin temel olarak dayandığı veriler bugün artık modern derlemlerdir. Klasik fişleme yönteminin hem emek yoğun bir süreci kapsaması hem de fişe yazılan bilgilerin tekrar elektronik ortama girilmesiyle oluşan tekrarlı süre alıcı çalışma nedeniyle bu yöntem terk edilmiştir. Modern ilişkisel veri tabanı mimarileriyle sözlük verisinin makineye girilmesi kolaylıkla yapılmakta, geriye dönük olarak verinin sorgulanması da yine saniyeler içinde yapılmaktadır. Günümüzde dilsel veri tabanlarına yönelik kuramsal çalışmaların giderek arttığı da görülmektedir. Sözlükte yer alacak birimlerin sıklıklarının otomatik yöntemlerle hesaplanması ve işlenmesi de yine derlemler üzerinden DDİ uygulama yazılımlarının da yardımıyla yapılabilmektedir. Bu konuda üzerinde en çok çalışılan dil İngilizcedir. Özellikle İngiltere ve Amerikan üniversitelerinde özel olarak oluşturulmuş metin işleme birimleri bir yandan İnternete dayalı veriden bilgi çıkarımına yönelik yöntem bilgisel uygulamaya yönelik çalışmalar yapılırken bir yandan da sözlük bilimcilerin ihtiyacı olan dilsel veriler de güncel metinler izlenerek elde edilmektedir.

4. Dünyada Başlıca Büyük Boyutlu Derlemler

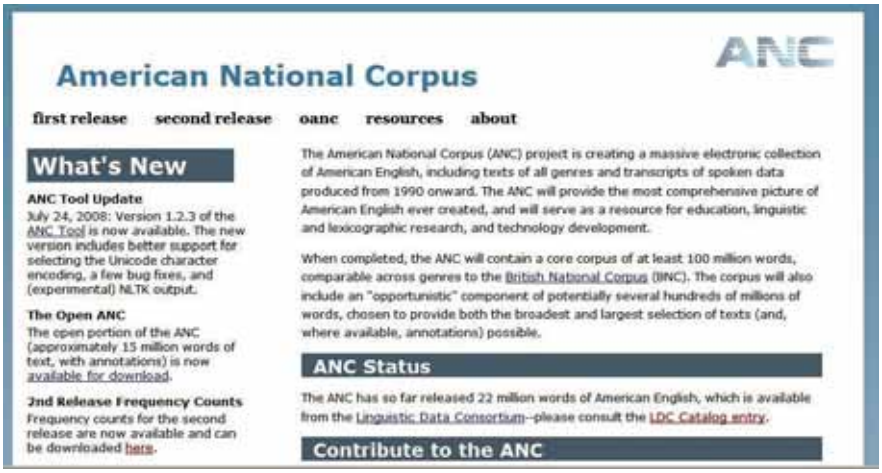
Elektronik biçimde, makinece okunur formatta ilk hazırlanan derlem 1 milyon sözcüklük işaretlenmiş bir derlem olan *Brown Derlemidir (The Brown Corpus)*. W. N. Francis ve H. Kucera'nın Brown Üniversitesi bünyesinde hazırladığı derlem 15 farklı metin türü dikkate alınarak hazırlanmış, günümüzde elektronik ortamda tutulan derlemlerin ilki olması dolayısıyla önemini korumaktadır. Boyutça küçük olmasına karşın günümüzde bu derlemden yararlanılmaya devam edilmektedir. 15 farklı metin türü (röportaj, tanıtmalar, din, hobi, resmi belgeler, kurgu, bilim, mizah gibi) altında 500 adet belge bir araya getirilmiş, her bir belge de yaklaşık 2000 sözcük derlenmiştir.

1991-1994 yılları arasında hazırlanan *İngiliz Ulusal Derlemi (British National Corpus)* 100 milyon sözcüklük bir derlemdir. Derlemin %90'ı yazılı metinlerden %10'u da konuşma dilinden verilerle oluşturulmuştur. 1994'ten sonra derleme yeni bilgi girişi yapılmamış ancak 2001'e kadar derlem geriye dönük olarak yeniden gözden geçirilerek BNC World adıyla ikinci sürümü hazırlanmıştır. Son olarak da 2007'de bir belge işaretleme standardı olan XML ile işaretlenmiş biçimi BNC XML Edition kullanıma sunulmuştur. Bu derlem de Brown Derlemi gibi değişik metin türlerinden (gazeteleri, dergiler, kurgu metinleri günlükler, mektuplar gibi) veriler derlenerek oluşturulmuştur. Derlemde toplam 3144 yazılı metne karşılık 910 adet konuşma diline yönelik belge bulunmaktadır. BNC için hazırlanmış web sayfalarıyla, derlemi sorgulamaya yönelik özgül yazılımlar da site üzerinden tanıtılmaktadır. TEI (Text Encoding Initiative) adı verilen derlem işaretleme standardıyla işaretlenen derlem için CLAWS adlı yazılım kullanılarak söz türü işaretleme otomatik olarak gerçekleştirilmiştir. Xaira ve SARA adlı iki yazılımla derlem ayrıntılı olarak sorgulanabilmektedir.

İngiliz Ulusal Derlemi web sitesi anasayfası

Birmingham Üniversitesinde hazırlanan *Bank of English* monitör derlem olarak adlandırılan ve sürekli veri güncellemesiyle derlenen dili mümkün olduğunca geniş bir şekilde modellemenin oluşturulmaya çalışıldığı derlem türünün bir örneğidir. 2005 yılı itibariyle yaklaşık 525 milyon sözcüğün elektronik ortamda tutulduğu derlemin 56 milyon sözcüklük bölümü dil öğretimine özel olarak oluşturulmuş alt derlemdir. Bu derlem için de BNC'de olduğu gibi internetten kullanıcılara erişim sağlama olanağı sağlanmıştır. İngiliz İngilizcesi, Amerikan ve Avustralya İngilizcesi veri dilidir.

BNC'ye paralel olarak Amerikan İngilizcesi temel alınarak geliştirilen ve halen çalışmaları süren *Amerikan Ulusal Derlemi* (*American National Corpus*), bugüne kadar 22 milyon işaretlenmiş sözcük içermektedir. Derlemin sözcük boyutu 100 milyon olarak hedeflenmektedir. Derlem için ayrıntılı listeleme ve çeşitli araçlar internet sitesinden sağlanabilmektedir. Derlem içeriğine yönelik sayısal bilgilerin de yer aldığı sitede Open ANC adıyla ücretsiz olarak derlemin bölümlerine ulaşılmaktadır.



Amerikan Ulusal Derlemi web sitesi anasayfası

5. Türkçede Derlem Çalışmaları

Derlem dil biliminin uluslar arası bilim çevrelerindeki öneminin internetle birlikte arttığı söylenebilir. İnternette basit bir arama motorundan elde edilen sorgulama sonuçlarının dünyanın bilgisinin metinler ve dil üzerinden sağlandığı dikkate alındığında, bilgisayar destekli dil bilimi çalışmalarının da derlemler yoluyla güçlendiği anlaşılmıştır. Özellikle 90'lı yıllardan bu yana gerek derlem çalışmalarının yöntem bilgisi gerekse derlem yazılımlarının kullanımıyla ilgili yayın sayısında önemli ölçüde artış yaşanmıştır.

Türkçede derlem kuramıyla ilgili olarak henüz bir kitap bulunmamakla birlikte TDK'nin gerçekleştirdiği iki bilgisayar destekli dil bilimi çalıştayında konuyla ilgili bildiriler bulunmaktadır.

Geleneksel dil çalışmalarında metne dayalı çalışmalar öteden beri yapılmakla birlikte, tek tek metinlerden değil de metin bütünlerinden topluca yararlanmak, bütünler arası karşılaştırma yapmak gibi derlem çalışmalarında ayrı ilgi ve çalışma alanı konuları üzerine kapsamlı olarak çalışmalar Türkiye'de yeni başlamıştır. Say (2006), dil bilimciler arasında yapılan bir ankette, araştırmacıların % 41'inin çalışmalarının derleme dayalı olduğunun ortaya çıktığını belirtmektedir. Derlem olarak nitelenen bu çalışmaların büyük çoğunluğu, kişisel çabalarla elektronik formlar dışında toplanmış ve paylaşıma açık olmayan yapıdadır.

Türkçe DDİ uygulama çalışmalarına kaynaklık edebilecek dengeli ve temsil gücü hesaplanmış ilk derlem Orta Doğu Teknik Üniversitesinde (ODTÜ) hazırlanan ve 1999-2003 yılları arasında tamamlanan ODTÜ Türkçe Derlemi'dir. Derlem, 1990 sonrasında toplanan yazılı kaynaklara dayalı metinlerden oluşmaktadır. Derlem genel amaçlı bir derlem olarak nitelenmiş, sözlü dile dayanan örnekler bir araya

getirilmemiştir. Derlem dil bilimi literatüründe denge ve temsil gücüne yönelik hazırlama ölçütleri de bu derlemin hazırlanmasında göz önüne alınmış, böylece 14 metin türü her biri 2000 sözcüklük 999 örneklem (201 kitap, 87 dergi 3 günlük gazete) bir araya getirilerek toplam 2 milyon etiketlenmiş sözcüklük derlem oluşturulmuştur.

Derlem için metin toplama, işaretleme ve sorgulamaya yönelik yazılımın geliştirilmesi çalışma planı izlenmiştir. Metinler Unicode salt metin olarak kaydedilmiş, derlemde yer alan sözcükler için Oflazer'in 1994'te hazırladığı biçim birimsel çözümleyici kullanılmıştır. XEROX Araştırma Merkezi Sonlu Durum Araçları kural tabanlı çözümleyici olarak biçim birimsel çözümlemede kullanışmıştır. Derlem daha sonra Sabancı Üniversitesiyle ortaklaşa yürütülen bir proje kapsamında ODTÜ-Sabancı Ağaç Yapılı Derlemi adını almıştır. Bu ortak proje kapsamında 7. 300 cümle ve 65.000 sözcük etiketlenmiştir.

Biçim birimsel çözümlemede birden çok sonuç veren yapılar için insan kontrolüne dayalı bir etiketleme standardı oluşturulan işaretleme yazılımında yer alarak belirlenen XCES XML tabanlı derlem etiketleme standardı yönergesi izlenerek etiketlemenin son olarak varsa hatalarının düzeltilmesi etiketleyici kişiye bırakılmıştır. Dil bilimciler için sorgulama yazılımında *düzenli deyimler* (regular expressions) ve mantıksal sorgulama (ve, veya) seçenekleri oluşturulmuştur. Biçim birimsel çözümleme dışında cümle düzeyinde otomatik çözümleme için de *bağlılık grameri* (dependency grammar) adı verilen ve Türkçe için çekim öbeklerinin belirlendiği bir form göz önüne alınmıştır (Say, 2006: 85). Derlem ve yazılımının İnternet ortamında araştırmacılara ücretsiz olarak paylaşıma açık biçimi bulunmaktadır.

Türkçe için derlem geliştirme çalışmalarında Çebi ve Varlıkların (2006) oluşturduğu ve TurCo adı verilen işaretlenmemiş derlemdir. TurCo'nun İnternet üzerinden toplanan bölümü 10 farklı web sitesinin İnternet tarayıcı yazılımının salt metin kaydetme özelliği kullanılarak oluşturulmuştur. Bu işlemde en büyük sorun olarak site sayfalarında geçen bağlantı bilgileri (hyperlink), tekrarlanan menü adları gösterilmiştir (Çebi ve Varlıklar 2006: 60).

Derlemdeki toplam sözcük sayısı50.111.828 olarak verilmiştir. Farklı sözcük sayısı 686.804 olarak çıkarılmıştır. Derlemin %90,40'ı internette yayımlanan gazete, dergi gibi güncel yayınlar ve Türkiye Büyük Millet Meclisi'nin internet sitesinde yayımlanmış metinlerden ve Devlet İstatistik Enstitüsü'nün web sayfaları taranarak elde edilmiştir. Derlemin bilgisayar ortamındaki toplam büyüklüğü 362 MB'dir. TBMM tutanaklarının yazılı ve sözlü Türkçeyi temsil ettiği derlem bilgisi olarak belirtilmektedir.

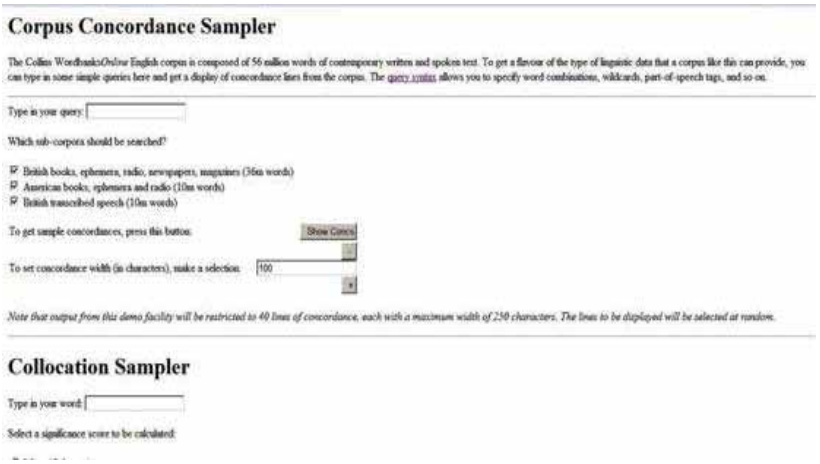
6. Günümüz İnternet Ortamında Derlem Dil Bilimi

Matbaanın bulunmasıyla hızlanan yazıya dayalı kültür ve getirdiği değişim bugün internet ortamında bulunan ve çevrimde dolaşan veri yoğunluğuyla daha da hızlanmıştır. İnternete dayalı öğretimden her türlü bankacılık işlemlerine ve ticarete varıncaya kadar hemen bütün veriler dilsel bir nitelik taşımaktadır. Crystal'ın (2001) da dile getirdiği gibi internet ne durumda olursa olsun sağlayacağı veri ve temeli

yazıya ve metne dayanacaktır (Yıldırım ve Tahiroğlu 2006:363).

İnternetin sağladığı metin ortamı HTML olarak bilinen bir işaretleme diline dayanmaktadır. Özellikle sürekli güncellenen günlük gazetelerin İnternet sayfaları ir veri tabanına bağlı html çıktılarından oluşmaktadır. Bununla birlikte akademik yazı biçimlerinin daha çok PDF ve MS Word formatında sunulduğu dikkati çekmektedir. Bu tür ortamlardan verinin ham olarak elde edilmesinde kullanılan ve kullanıcının İnternette çevrimdışı gezinti yapmasına olanak sağlayan ticari ve açık kaynak kodlu yazılımlar bulunmaktadır. Bu yazılımlarla istenilen türdeki sayfa biçimleri istenilen zamanda ve sıklıkta bir web sitesindeki metni kopyalayabilmektedir. İnternet üzerinden metin arşivi oluşturmada bu yazılımlar sıklıkla kullanılmaktadır.

BNC, ANC gibi büyük boyutlu işaretlenmiş derlemlerin bir form aracılığıyla sorgularını sağlayan veri tabanları yayımlanmaktadır. Collins Cobuild Concordance and Collocations Sampler adlı çevrimiçi yazılım, İngilizce hazırlanan ve 56 milyon etiketlenmiş sözcüğün standardı kendi içinde oluşturulmuş etiketleme sistemiyle sorgulanmasına olanak vermektedir.



Cobuild Concordance and Collocations Sampler çevrim içi derlem sorgulayıcısı.

İstenilen bir web sayfasından sözcük sorgulama, bağlamlı dizin oluşturma, mantıksal sorgulama gibi işlevleri bulunan WebAsCorpus adlı çevrim içi yazılım internetteki güncel veriyi derlem olarak kullanmaktadır. Türkçe için de sorgulamalara olanak tanıyan yazılım basit ve gelişmiş sorgulamalara en fazla sorgulanacak web sayfası, sözcüğün bulunduğu harf bağlamının seçilebildiği ayrıntıları da barındırmaktadır.

WebAsCorpus bağlamlı dizin oluşturucusu anasayfası.

7. Sonuç

Bilgi teknolojileri, derlem dil bilimini özellikle bilgisayar bilimleriyle disiplinler arası bir noktaya götürmektedir. Bu bağlamda derleme yönelik kuramsal ve uygulamaya yönelik akademik disiplinlerin Türkçe çalışmalarında da kullanılması gerektiği açıktır.

Dilsel değişmelerin anında ve güncel olarak takip edilebileceği İnternet derlemlerinin Türkçe dil bilgisi yazarlığının yanında özellikle sözlük bilimine katkısı büyük olacaktır.

Kaynakça

Baker, Paul, Andrew Hardie, Tony McEnery (2006), *A Glossary of Corpus Linguistics*, Edinburgh University Pres Ltd, 22 George Squarei Edinburgh.

Çebi, Yalçın- Özlem Varlıklar (2006), "Türkçe Derlem Oluşturmada Karşılaşılan Sorunlar ve Çözüm Önerileri", *Bilgisayar Destekli Dil Bilimi Çalıştayı Bildirileri*, TDK Yay: 868, Ankara

Kennedy, Graeme (1998), *An Introduction to Corpus Linguistics*, Addison Wesley Longman Limited, England.

McEnery, Tony, Andrew Wilson (2004), *Corpus Linguistics*, Edinburgh University Pres.

Say, Bilge (2006), "Türkçe İçin Bir Derlem Geliştirme Çalışması", *Bilgisayar Destekli Dil Bilimi Çalıştayı Bildirileri*, TDK Yay: 868, Ankara

Yıldırım, Faruk, B. Tahir Tahiroğlu (2006), "İnternette Türkçe Kullanımı Sorunu", *Türkçenin Çağdaş Sorunları* (Hazırlayanlar Gürer Gülsevin, Erdoğan Boz), 2. baskı, Gazi Kitabevi.

Erişim

http://www.natcorp.ox.ac.uk/corpus/index.xml

http://www.essex.ac.uk/linguistics/clmt/w3c/corpus_ling/content/introduction.html

http://corpora.ids-mannheim.de/ccdb/

http://webascorpus.org/searchwac.html

http://www.collins.co.uk/Corpus/CorpusSearch.aspx