
Low-resource Languages Toolkit

Yayım 0.1

Alp Öktem, Col·lectivaT

20 Eki 2022

1	Giriş	3
1.1	Bu belge kimler için hazırlandı?	3
1.2	Katkıda bulunabilir miyim?	4
1.3	Yazarlar	4
2	Diller ve Dijital Çağ	5
2.1	Dijital yok olma	5
2.2	Bir dilin çevrimiçi varlığı	6
2.3	Dil dokümantasyon projeleri	7
2.4	“Geleneksel” teknoloji	7
3	Veriye dayalı dil teknolojileri	17
3.1	Makine çevirisi (<i>machine translation</i>)	18
3.2	Otomatik konuşma tanıma (<i>automatic speech recognition</i>)	19
3.3	Metin-konuşma sentezi	20
4	Dil verileri	23
4.1	Metin derlemi	23
4.2	Paralel veri (bitext)	24
4.3	Konuşma derlemi	26
5	Örnek çalışmalar	29
5.1	Judeoespanyol: Akdeniz’in iki yakasını birleştiren dil	29
5.2	Tabandan örgütlenen NLP toplulukları	32
5.3	Common Voice kampanyaları	33
5.4	Diğer girişimler	33
6	İyi uygulamalar	35
7	Lisans	37



Şekil 1: Bu proje Avrupa Birliği tarafından finanse edilmiştir.

Bu belgede az kaynaklı ve yok olma tehlikesi altında olan dillerin korunmasında teknolojinin, özellikle yapay-zeka bazlı teknolojilerin rolünü ele alıyoruz.

Bu belge, Col-lectivaT Kooperatifi ve Sefarad Kültürü Araştırma Derneği (SKAD) tarafından yürütülen “Judeoespanyol: Akdeniz’in iki yakasını birleştiren dil” projesi kapsamında hazırlanmış yaşayan bir belgedir. Bu proje, T.C. Kültür ve Turizm Bakanlığı tarafından Avrupa Birliği’nin mali desteği ile hayata geçirilen “Ortak Kültür Mirası: Türkiye ve AB Arasında Koruma ve Diyalog-II (CCH-II) Hibe Programı” kapsamında hibe desteği almıştır.

Not: This document is also available in Turkish and Spanish.

[Bu doküman Türkçe de erişilebilir.](#)

[Este documento también está disponible en español.](#)



Şekil 2: Bu proje Avrupa Birliği tarafından finanse edilmiştir.

Şu anda, yeterli kaynağa sahip diller ile daha az kaynağa sahip diller arasında gittikçe büyüyen bir dijital uçurum var. Bu uçurum, az kaynaklı diller için dijital yok olma tehlikesini daha da artırıyor. Çoğunluk dilleri için, web’de büyük oranda var olmaları nedeniyle, faydalı araçlar ve kaynaklar oluşturma süreci çok daha kolay. Ancak, birçok azınlık dili, bu tür araçların yaratılmasını sağlamak için yeterli maddi ve insan kaynağına sahip değil. Devlet desteğinin olmaması, kamusal görünürlük, toplumsal ve kurumsal baskı, bu dillere günümüzün dijital alanlarında öncelik verilmemesinin doğrudan nedenleridir.

Dillerin korunmasına yönelik çabalar, esas olarak dil dokümantasyonu, öğretimi ve fiziksel topluluk oluşturmaya odaklanır. Göz ardı edilen bir alan ise yapay zekaya dayalı araçların oluşturulmasıdır. Makine çevirisi, konuşma sentezi ve konuşma tanıma gibi araçlar, artık insan-makine arayüzleri oluşturmada önemli eşdeğerlerdir. Ayrıca bu araçlar, ölmekte olan dillerin dil birikimini modellemeye ve bu dilleri gelecek nesiller için korumaya yardımcı olabilir.

1.1 Bu belge kimler için hazırlandı?

Bu belge:

- Kendi dillerindeki araçları ve kaynakları genişletmekle ilgilenen *dil aktivistleri*
- Araştırma ve dil teknolojilerinin yaratılması için veri toplamakla ilgilenen *dilbilimciler*
- Çalıştıkları dil için verileri artırmakla ilgilenen *doğal dil işleme (NLP) araştırmacıları*
- Kaynakları yetersiz yerel ve küresel dillerin yeniden canlandırılmasına katkıda bulunmak isteyen *dil aktivisti destekçileri* için hazırlanmıştır.

1.2 Katkıda bulunabilir miyim?

Bu belge, açık lisanslı (CC-BY) yaşayan bir belgedir. Kaynak dosyası <https://github.com/CollectivaT-dev/language-toolkit> adresinde herkese açık olarak paylaşılır. Buradan kendiniz çalışmak için bir sürüm çekebilir ve ardından katkınızı gönderebilirsiniz. Bunlar, yazım hatalarını düzeltmek, çeviri eklemek, bir bölümü detaylandırmak ve kendi çalışmanızı açıklamak gibi katkılar olabilir. Tereddütlerinizi info@collectivat.cat adresinden bize iletebilirsiniz.

1.3 Yazarlar

- Alp Öktem - (Web, Twitter)
-



Şekil 1: Bu yayın Avrupa Birliği'nin maddi desteği ile hazırlanmıştır. İçerik tamamıyla Col-lectivaT ve SKAD'ın sorumluluğundadır ve Avrupa Birliği'nin görüşlerini yansıtmak zorunda değildir.



Şekil 2: Bu proje Avrupa Birliği tarafından finanse edilmiştir.

Diller ve Dijital Çağ

Dijital çağ birçok yeni fırsat ve avantaj getirdi. Hiç şüphesiz, insanları kısa sürede hayal bile edilemeyecek şekillerde birbirine bağladı. Bununla birlikte, her yenilik bazı zorlukları ve tehditleri beraberinde getirir. *World Wide Web* (nere-deyse) herkesin erişebildiği bir kaynaktır, ama sadece bir avuç dil tarafından yönetilmektedir.

Örneğin İngilizce, dünyanın sadece %15'i tarafından konuşulmasına rağmen Web'deki tüm içeriğin %54'ünü elinde tutmaktadır. Öte yandan, jeopolitik hakimiyetlerine rağmen Rusça, Çince ve İspanyolca gibi dillerin her biri Web'deki içeriğin yaklaşık %5 ila %6'sını temsil ediyor.

Bu, Dünya'nın yok olma tehlikesiyle karşı karşıya ve ölmekte olan birçok dilini nereye koyuyor? Ne yazık ki, bu resmin en köşelerine.

Diller, toplumlar içinde gelişir ve günlük kullanımla nesillere aktarılır. Hayatlarımız dijital ortamlar aracılığıyla ne kadar bağlıysa, çevrimiçi (*online*) olarak temsil edilmeyen anadillerimizle giderek daha az iletişim kuruyoruz. Bu da, son raddede bu dillerin genç nesiller tarafından daha az kullanılmasına yol açıyor.

Not: Nasıl? Örneğin, Türkiye'de Kürtçe konuşan bir kişi, sağlık hizmetleri web sitesine giriyor ve her şeyin Türkçe olduğunu görüyor, veya en son çıkan online sosyal medya platformuna bakıyor ve varsayılan dil olarak İngilizce olduğunu görüyor. Karşılaşılan bu küçük durumlar, yollarını bulmak ve ihtiyaçlarını gidermek için anadillerini konuşmanın yeterli olmadığını düşündürüyor. Çoğu zaman size en yakın çoğunluk dilini, hatta birçok kez başka dilleri de bilmek zorunda olduğunuzu hissettiriyor.

2.1 Dijital yok olma

UNESCO'ya göre, bu yüzyılın sonuna kadar yaklaşık 3.500 dilin yok olması bekleniyor ve bunda teknolojinin rolünü inkar edemeyiz. *Kornai*, dillerin büyük çoğunluğunun (%95'in üzerinde) dijital olarak yükselme kapasitesini hali hazırda kaybettiğini belirtiyor. Dijital yükselme bir dilin, bir Wikipedia sayfasının o dilde yönetiminden dil derslerinin erişime açılmasına veya dil teknolojisi verilerinin oluşturulmasına kadar çok çeşitli dijital bağlamlarda kullanımını gerektiriyor.

Elbette, tek başına teknolojiyi dil kaybının günah keçisi ilan etme hatasına düşmemeliyiz. Toplumda zaten var olan güç dinamiklerinin bir temsili sadece. Belirli dilleri önemsiz kılan hatta baskı altına alan devletler, dijital dönüşümlerine tüm

bu dilleri dijital altyapılarının dışında bırakarak başlıyorlar. Günümüzde, Kuzey Amerika ve Avrupa merkezli büyük teknoloji şirketleri, varsayılan dil olarak önce İngiliz yaklaşımını izliyor.

Teknoloji ayrıca dil koruma, bilgi paylaşımı ve dil dokümantasyonu etrafında topluluklar oluşturmak için de kullanılabilir.

2.2 Bir dilin çevrimiçi varlığı

Geleneksel olarak, bir dil aktivistinin sorumlulukları, dili aktif olarak konuşmak, genç nesillere aktarmak, dil öğrenme ve konuşma toplulukları oluşturmak, dilinin dahil edilmesi için kamu kurumlarıyla müzakere etmek, dilinin dokümantasyonu için dilbilimcilerle işbirliği yapmak gibi sayılabilir.

Günümüzde zorluk, dili sadece fiziksel dünyada değil, aynı zamanda çevrimiçi ortamda da canlı kılmaktır. Bu, bir dilin hayatta kalmasıyla iki şekilde ilintilidir:

1. Karşılıklı bilgi alışverişi ve çevrimiçi görünürlük, bir dili öğrenmekte olan veya öğrenmeye yeni başlayan kişilerin ilgisini çeker ve bu ilginin sürmesini sağlar.
2. Çevrimiçi olarak depolanan şey, belgeleme ve teknoloji geliştirmeye yardımcı olan dijital bir dil kayıdır.

Aşağıda, internetin hangi yollarla hem çok dilli ve çoğul hale gelip hem de yok olma tehlikesiyle karşı karşıya olan dillerin canlanmasına yardımcı olduğunu açıklayacağız.

2.2.1 Bilgiye erişim

Dilleri çevrimiçi hale getirme konusundaki en popüler girişimlerden biri Wikipedia'dır. Wikipedia, açık bir ortaklaşa çalışma ve gözden geçirme sistemi aracılığıyla bir gönüllüler topluluğu tarafından yazılan ve sürdürülen açık ve özgür bir çevrimiçi ansiklopedidir.

Wikipedia'nın geniş anlamıyla amacı bilgiye erişimi demokratikleştirmektir. Doğal olarak, bu ethos etrafında inşa edilen kültür, çok dillilik ile paralel gider. Sadece İngilizce ile başlamasına rağmen, hızla dünyanın birçok diline yayılmıştır. 326 farklı dilde bulunan (03.05.22 itibarıyla) ve dil sayısı gittikçe artan Wikipedia'nın, dil açısından internetteki en çeşitli platform olduğunu söylemek büyük ihtimalle yanlış bir ifade olmaz.

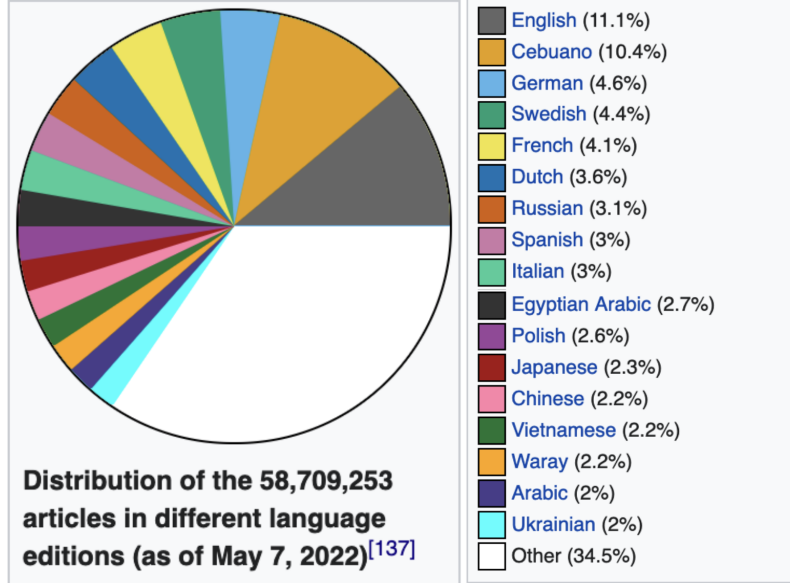
Not: İngilizce olmayan bir dilde ilk Wikipedia düzenlemesi 16 Mart 2001'de Katalanca yapıldı. Bugün, bir azınlık dili olmasına rağmen çok sayıda kaliteli makalesiyle büyük bir çevrimiçi varlık gösteren [Katalanca Vikipedi'nin](#) 20. büyük wikipedia olması dikkate değer.

Wikipedia'da yeni bir dili kullanıma sunmak kolay bir iş değil ([bkz.](#)). Ancak bilgiyi çevrimiçi ortamda erişilebilir kılmak ve onun etrafında sanal bir topluluk oluşturmak için kesinlikle çok iyi bir yol.

2.2.2 Dilin yoğun kullanımı

Bir dilin teknoloji yardımıyla yeniden canlandırılmasına en güzel örneklerden biri Yidiş'tir. Yahudi Soykırımından sonra Yidiş konuşanların sayısı önemli ölçüde azaldı (yaklaşık 10 milyon konuşmacıdan). Hayatta kalanlar, zulümden kaçmak için yerleştikleri topraklarının dilini özümsemek zorunda kaldılar. Geçen yüzyılda, küçük ve dağınık halde bulunan Hasidik topluluklar dışında Yidiş kullanımı neredeyse ortadan kalkmıştı.

İnternetin yükselişi ve çevrimiçi forumların popülerleşmesi ile birlikte, Yidiş konuşanlar bu platformları kendi dillerinde sohbet etmek için kullandılar. Zamanla sanal dünya Yidiş konuşanlar için, Idishe Velt (Yahudi Dünyası) ve Kave Shtiebel (Kahve Evi) gibi forumlarda başlıca buluşma noktası haline geldi.



Şekil 1: Farklı Wikipedia'ların birbirilerine göre büyüklükleri (kaynak)

2.3 Dil dokümantasyon projeleri

Geçen yüzyılda dillerin hızla kaybolması, dil belgeleme ve canlandırma için çalışan birçok girişimi güçlendirdi. Bu girişimlerden biri, yok olma tehlikesiyle karşı karşıya olan dilleri güçlendirmeye yardımcı olmak için, dil meraklıları, dilbilimciler ve endüstri ortaklarından oluşan bir işbirliği merkezi olarak hareket eden web tabanlı platform *The Endangered Languages Project*'tir. Web sitesinin kullanıcıları, kolay aramayı mümkün kılan özgün bir coğrafi etiketleme sistemi kullanarak dil örneklerini metin, ses, bağlantı veya video formatında yükleyerek katkıda bulunabilirler.

Benzer şekilde, 2014 yılında başlayan *Wikitongues*, dünyadaki dillerin kayıtlarını ve kaynaklarını toplamaktadır. Şu anda 700'den fazla dilde videolar, 200 dilde leksikon ve yüzlerce harici kaynağa bağlantılar içerir.

2.4 “Geleneksel” teknoloji

Bir dili korumak, yalnızca kelimeleri veya cümleleri kaydetmek ve onları çevrimiçi bir kasada saklanmak üzere dijitalleştirmek değildir. Dil, doğası gereği insanlarla, kültürle ve kimlikle ilgilidir. Bir dili canlı tutmak için, o dilin birçok kişi tarafından konuşulması, günlük kültürün içine girmesi ve gelecek nesillere aktif olarak aktarılması gerekir. Günümüzde internet, sosyal medya, yazılımlar ve platformlar günlük hayatımızda büyük bir yer kaplamaktadır. Bu bölümde, bir teknolojinin dijital olarak gelişmesi için gereken temel araçlardan bazılarını listeleyeceğiz.

2.4.1 Unicode destekli yazı tipi

Dijital yazı tipi, bilgisayarların karakterleri sizin dilinizde nasıl görüntüleyeceğini bilmelerinin yoludur. Unicode (Evrensel Kod), resmi adıyla *Unicode Standard*, dünyadaki çoğu yazım sistemlerinden metinlerin tutarlı bir şekilde kodlanması, gösterilmesi ve işlenmesi için oluşturulmuş bir bilgi teknolojisi standardıdır. Unicode Konsorsiyumu tarafından sürdürülen standart, 159 modern ve tarihi yazım sisteminin yanı sıra sembolleri, emoji'leri ve görsel olmayan kontrol ve biçimlendirme kodlarını içeren 144.697 karakteri tanımlar.

Google Noto'ya giderek ve 500'den fazla yazı sistemini temsil eden yazı tipleri arasında arama yaparak dilinizin bir yazı tipi tarafından desteklenip desteklenmediğini kontrol edebilirsiniz. Eğer desteklenmiyorsa, bir yazı tipi tasarımcısının yardımıyla kendi diliniz için yazı tipi oluşturabilir ve manuel olarak bilgisayarınıza yükleyebilirsiniz.



Endangered Languages Project

Supporting and celebrating global linguistic diversity

[Map](#) | [Languages](#) | [Resources](#) | [Submit](#) | [Blog](#) | [Download](#) | [About](#)

Ladino

[aka *Judeo-Spanish*, *Sephardic*, *Hakitia*]

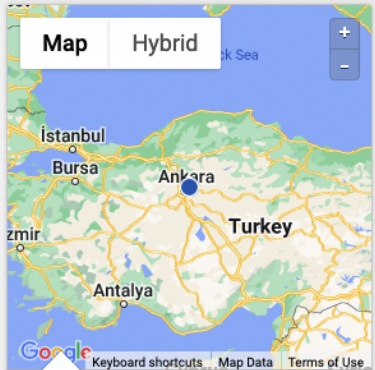
Classification: Indo-European * at risk

[Description](#) | [Resources](#) | [Activity](#) | [Revitalization](#) | [Bibliography](#) | [Suggest a Change](#) | [Subscribe](#)

Language metadata

The name Ladino refers most commonly to the written/literary form of the language. Most speakers refer to the spoken language as Judeo-Spanish.

ALSO KNOWN AS	Judeo-Spanish, Sephardic, Hakitia, Haketia, Judeo Spanish, Sefardi, Dzhudezmo, Judezmo, Spanyol, Haquetiya
CLASSIFICATION	Indo-European, Italic, Romance, Western Romance
CODE AUTHORITY	ISO 639-3
LANGUAGE CODE	lad
VARIANTS & DIALECTS	• Ladino • Judezmo • Haquetiya
DOWNLOAD	As csv
MORE RESOURCES	OLAC search



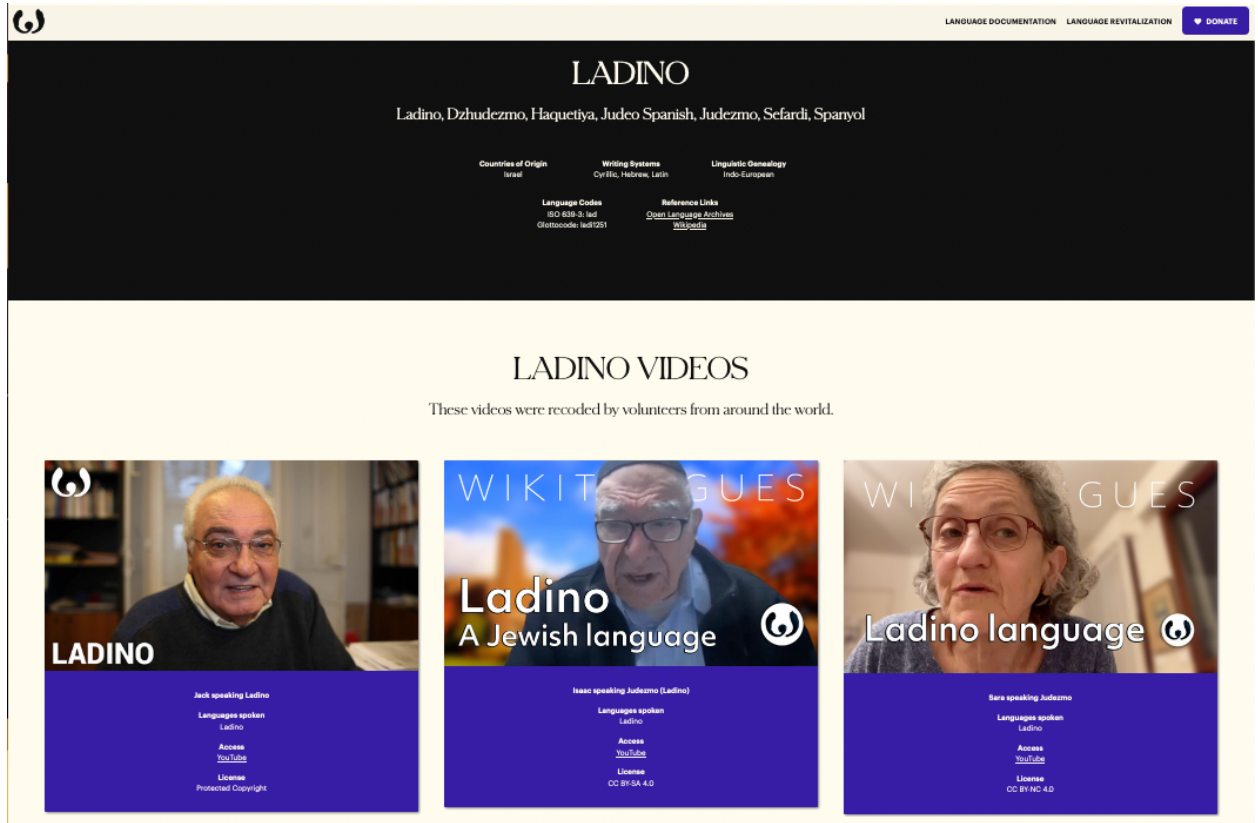
Map Hybrid

Location information: COORDINATES 40.0,33.0

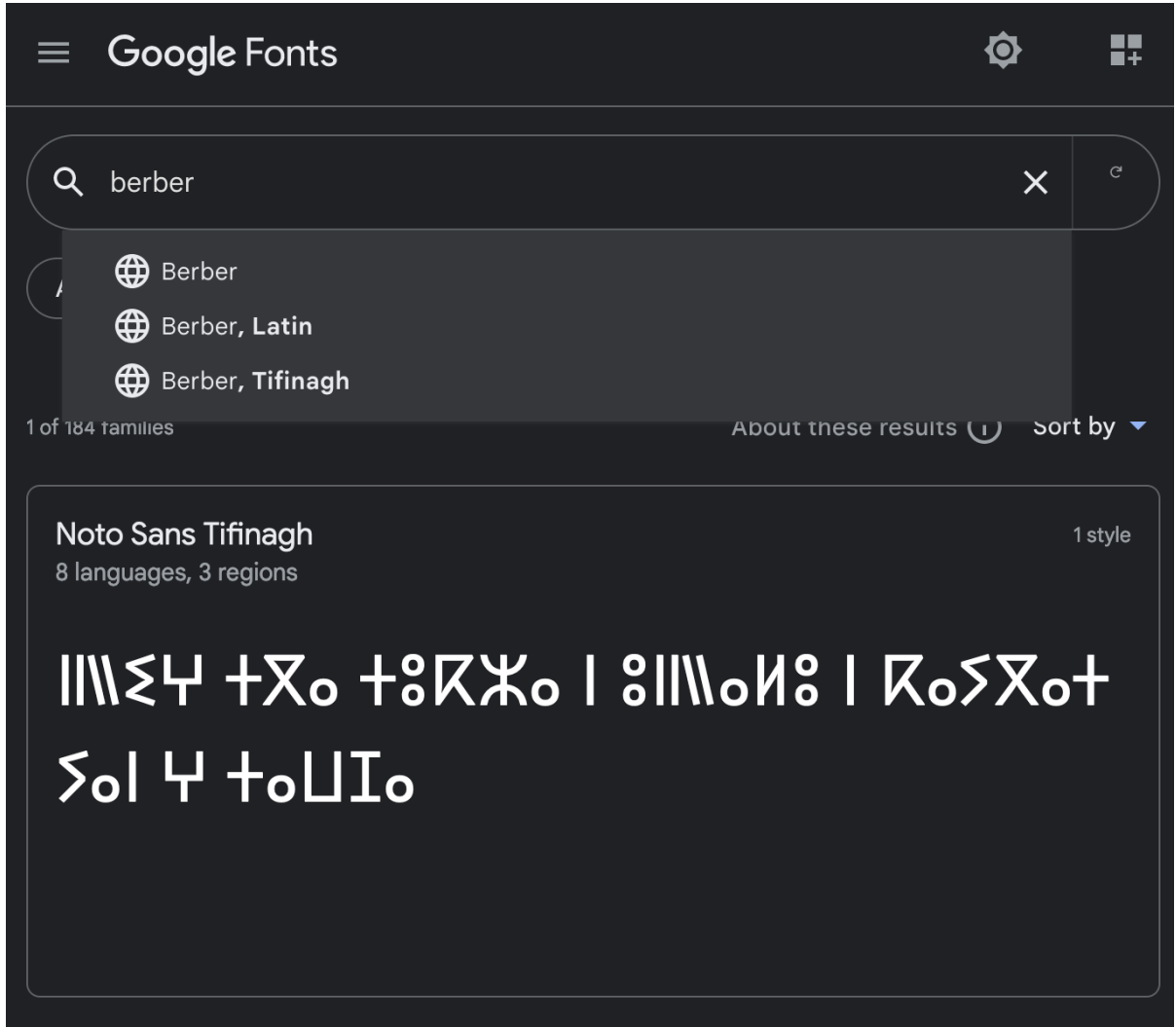
COMPARE SOURCES (2)

Information from: "The World Atlas of Language Structures". Bernard Comrie and David Gil and Martin Haspelmath and Matthew S. Dryer · Oxford University Press

Şekil 2: *Endangered Languages Project*'te Ladino



Şekil 3: Wikitongues’da Ladino videolar



Şekil 4: Google Fonts'ta Fas Berberi Tifinag alfabesi

2.4.2 Klavye / Tuş takımı

Bilgisayarlarla doğal olarak konuşacağımız günler gelinceye kadar, onlarla iletişim kurmak için kullanacağımız en temel arayüz klavye olacaktır. Dünyadaki birçok dil için bulunmasına kolayca kesin gözüyle bakılan bir teknolojidir, ancak ne yazık ki dünyadaki tüm yazı sistemlerinde mevcut değildir. Eğer bir dil için geliştirilmiş bir klavye yoksa ya da yeterince gelişmemişse, o dili konuşanlar iletişim kurmak için diğer alfabeleri ve hatta dilleri tercih etme eğilimindedir. Örneğin, Amharca, Tigrinya ve Oromca gibi Etiyopya dillerini konuşanlar, akıllı telefonlarında Ge'ez alfabesi önceden yüklü olmadığı için, İngilizce kullanmaya geçerler. Birçok ülkede, Arapça konuşan gençler, eski mobil cihazlarda ve web teknolojilerinde Arap alfabesinin eksikliğini kapatmak için, Latin alfabesi karakterlerinden ve rakamlardan oluşan kendi *chat*-alfabeleri [Arabizi](#)'yi icat ettiler.

Telefonunuzda veya örneğin bilgisayarınızda size gerekli bir klavye mevcut değilse, bir klavyeyi aramanız veya kendi klavyesini oluşturmanıza yardımcı olacak bazı kaynaklar şunlar:

- Google'ın mobil klavyesi, [Gboard](#)
- 2000 dili destekleyen [Keyman](#)
- Microsoft Klavye Düzeni Oluşturucu (MSKLC)
- macOS için Unicode Klavye Düzeni Biçimleyicisi [Ukelele](#)

2.4.3 Çevrimiçi sözlük

Sözlük (leksikon), bir dili belgelemenin önemli bir yoludur, çünkü kelimeler ve anlamları için bir referans görevi görür. İnternet bağlantısı olan herhangi bir cihazdan erişilebilir olduğu için, bir çevrimiçi sözlüğün baskısı asla tükenmez. Ayrıca, açık kaynak sözlükler, hem dili konuşanları, hem de dilbilimcileri ve teknoloji uzmanlarını içeren bir topluluğun çabasıyla ortaklaşa biçimde yaşatılabilir ve büyütülebilir.

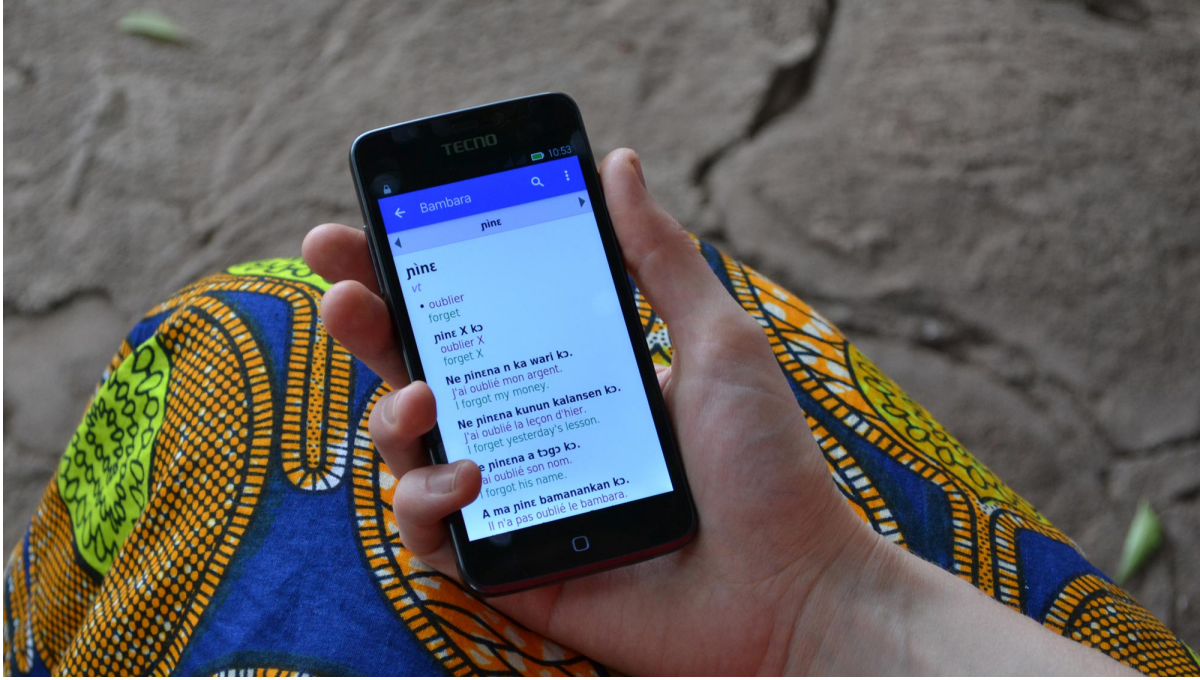
[Living Dictionaries](#) (Yaşayan Sözlükler), *Living Tongues Institute for Endangered Languages* (Tehlike Altındaki Diller için Yaşayan Diller Enstitüsü) tarafından oluşturulmuş bir çevrimiçi sözlük oluşturma platformudur. Dil topluluklarına dillerini koruma ve canlandırma çabalarında yardımcı olan kapsamlı, ücretsiz çevrimiçi teknoloji araçları sağlar. Ayrıca kelimelerin ve cümlelerin kaydedilmesine izin verir. Mayıs 2022 itibarıyla 237 dili desteklemektedir. *Living Dictionaries*'de dilinizin sözlüğünü oluşturmaya başlamak için, [Tanımlama listeleri](#)'nden yararlanabilir ve [YouTube kanallarındaki](#) eğitim videolarını izleyebilirsiniz.

[SIL Dictionary App Builder](#) “Android ve iOS akıllı telefonlar ve tabletler için özelleştirilmiş sözlük uygulamaları oluşturmanıza yardımcı olur. Kullanılacak sözlük veri dosyasını, uygulama adını, yazı tiplerini, renkleri, ‘hakkında’ bilgilerini, sesi, resimleri ve simgeleri siz belirlersiniz. Dictionary App Builder, her şeyi bir araya toplayıp sizin için özelleştirilmiş uygulamayı oluşturacaktır. Ardından uygulamayı telefonunuza yükleyebilir, Bluetooth ile başkalarına gönderebilir, microSD hafıza kartlarında paylaşabilir ve internetteki uygulama mağazalarında yayınlatabilirsiniz.”

2.4.4 Dil öğrenme uygulamaları

Çevrimiçi eğitim platformları, bugün birçok insanın dil öğrenmeye yaklaşım biçiminde devrim yarattı. Öğretmenin yerini tutmasalar da, bu platformlar ya geleneksel dersleri tamamlayıcı niteliktedir ya da bazı diller için tek seçenektir. Ayrıca, insanların herhangi bir cihazda (mobil veya masaüstü), kendi ilerleme hızlarında ve planlamalarında öğrenmelerini mümkün kılmak gibi birçok avantaj da sağlarlar. Bu uygulamalar, kısa, eğlenceli, hafif ve hızlı geçilen bölümlerle dil dersleri ve alıştırmalar sunar, öğrencilerin ilerlemelerini takip etmelerine, hatta çevrimiçi topluluk ortamlarında dil öğretmenleri ile sohbet etmelerine veya tutmalarına olanak verir.

Dünyada tehlike altında olan, azınlıklaştırılmış ve/veya yetersiz kaynaklara sahip olan dillerin çoğu, henüz çevrimiçi dil kursları oluşturmak için önemli bir çevrimiçi varlığa ya da yeterli dil belgelerine sahip değil. Bununla birlikte, dil topluluklarından gelen baskılar ve dünya çapında yerli dillerini öğrenmeye yönelik artan duyarlılık sayesinde, bu uygulamaları geliştiren şirketlerin tehlike altında olan dillere ve azınlık dillerine yatırım yapma konusundaki ilgisi arttı.



Şekil 5: Cep telefonunda Bambara sözlüğü kullanan bir kadın (Görsel: SIL International)

Maorice, İskoçça, Hawaii, Keçuva, Navahoca ve Lakotaca gibi diller Duolingo, Babbel ve uTalk gibi tanınmış eğitim platformlarında yerlerini almaya başladı.

Bu platformları dört şekilde sınıflandırabiliriz:

- **Modül tabanlı:** Bu uygulamaları kullanmak, bir okulda veya kursta ders almaya benzer. Kullanıcılar, eğitimciler tarafından planlanmış ve modüllerden oluşan bir müfredatı izler. Öğrencilerin ilerlemelerini takip etmelerine, bildirim almalarına ve puan kazanmalarına olanak tanır. Dikkate değer bazı örnekler şunlardır: [Duolingo](#), [Babbel](#), [uTalk](#), [Master Any Language](#). Ne yazık ki dil topluluklarının tek başına bu platformlara yeni bir dil eklenmesine ve kurulmasına karar vermesi mümkün değildir. Ancak, bu platformlardan bazılarının oluşturduğu topluluklara katılmak ve bunlar aracılığıyla yeni modüllerin oluşturulması için “lobi” yapmak mümkündür. Bu platformların çoğunun kar amacı güden platformlar olduğu ve ekonomik anlamda dil topluluklarının çalışmalarının karşılanmadığı dikkate alınmalıdır.
- **Oyun tabanlı:** Bu uygulamalar öğrenciye “soru-cevap” eşleştirmeleri verir ve kişinin zaman içinde eşleştirmeyi ne kadar iyi ezberlediğini değerlendirebilir. Bu yöntem, görsel ve işitsel yöntemlerle dil öğrenenler tarafından çevrimiçi popüler seçeneklerden biridir. Ayrıca matematik ve fen gibi diğer birçok çalışma alanı için de kullanılır. [Memrise](#), yerli dilleri ve diğer yetersiz kaynaklara sahip azınlık dilleri için en çekici uygulamalardan biri olabilir. Dil öğrenimini optimize etmek için, bilgi kartları, aralıklı tekrarlar ve görsel araçlar gibi hafıza tekniklerini kullanan çevrimiçi bir eğitim platformudur. Memrise’da, 170’in üzerinde dilde içerik bulunur, bu sayı diğer dil öğrenme platformlarının çoğundan çok daha yüksektir. Site, takdire değer şekilde, özellikle dünyanın dört bir yanında konuşulan yerli dilleri ve lehçeleri için, o dillerin toplulukları tarafından oluşturulmuş etkileyici bir içerik seçkisine sahiptir: Yupik (bkz alttaki ilk ekran görüntüsü), Çeroki (bkz alttaki ikinci ekran görüntüsü), Algonkin, Supik, Çoktav, Grönland, İnuit, Lakota, Nahuatl, Yukatek Maya, Kiçe, Keçuva, Guarani, Aynu, Jeju dilleri, ve Avrupa, Afrika, Orta Doğu, Asya ve Pasifik’te konuşulan diğer birçok orta ölçekli diller. Memrise, “kendin yapçı” (*DIY*), demokratik ve tabandan gelen bir platform hissi veriyor. Memrise yenilikçi bir platform; çünkü: 1) kullanıcılar hali hazırdaki kursları ve bilgi kartı setlerini takip edebiliyor ve bir kursta ilerlerken kelimeleri ve cümleleri hatırlamaya yardımcı olmak için kendi animasyonlu araçlarını sorunsuz bir şekilde yükleyebiliyor; 2) site, tekrarlar, küçük testler, kısa videolar, eğlenceli görüntüler ve ileri seviye konuşmacılar tarafından hazırlanmış kayıtlar aracılığıyla kullanıcıların dil içeriğiyle bağlantı kurmaları için ilgi çekici bir yol sağlıyor; ve 3)



Şekil 6: Kri dilini öğrenmek için *Master Any Language* platformunda bir içerik

platform, bir topluluğun kullanıcılarının kendi dillerinde, kolayca ve başkalarının da kullanabileceği dil kursları oluşturmalarına olanak tanıyor. Buna benzer platformlar şunlar: [AnkiApp](#), [Language Drops](#), ve [MosaLingua](#).

MEMrise

Home Courses Groups Subscribe

Courses > Languages > Native American > Quechua

Quechua Vocabulary (South Peru, Bolivia)

Created by FROG123

Tweet

Get started now

Levels (3)

1 Ready to learn Vocabulary / Vocabulario 1-30

2 Ready to learn Vocabulary / Vocabulario 31-60

3 Ready to learn Vocabulary / Vocabulario 61-88

Leaderboard

Week Month All Time

1. benvdh 0

2. mria10156 0

Şekil 7: Memris'te topluluk tarafından oluşturulmuş Keçuva öğrenme alıştırmaları

- **Sohbet tabanlı:** Bu uygulamalar, öğrencilerin interaktif bir canlı sohbet aracılığıyla ilgilerini çeken dili konuşanlarla bağlantı kurmasına olanak tanır. Bu, öğrenciler için stressiz ve sosyal bir ortam sağlar. [HiNative](#) ve [HelloTalk](#) gibi bazı örnekler son zamanlarda özellikle Asya ülkelerinde popülerlik kazandı.
- **Çevrimiçi öğrenci-öğretmen platformları:** Klasik öğretmen-öğrenci ilişkisini tercih eden ancak yakınlarda bir öğretmene erişimi olmayan öğrenciler için, [iTalki](#) ve [Verbling](#) gibi platformlar çevrimiçi derslerin ayarlanmasına yardımcı olur. Bu aynı zamanda, öğretmenler için doğrudan gelir sağladığı için, doğrudan dil topluluğuna da katkıda bulunur.

2.4.5 Kaynaklar

- [Wikipedia'da Katalan Wikipedi](#)
- [How Technology Can Save Rapidly Dying Languages](#)
- [Wikipedia'da Unicode](#)
- [Dil sürdürülebilirliği araç seti](#)
- [Learn Language Online by Living Tongues](#)
- [Wikipedia'da bilgisayar destekli dil öğrenimi](#)



Şekil 8: Bu yayın Avrupa Birliğinin maddi desteği ile hazırlanmıştır. İçerik tamamıyla Col-lectivaT ve SKAD'ın sorumluluğundadır ve Avrupa Birliği'nin görüşlerini yansıtmak zorunda değildir.



Şekil 9: Bu proje Avrupa Birliği tarafından finanse edilmiştir.

Veriye dayalı dil teknolojileri

Dijital devrim kapımızda ve Yapay Zeka (AI) konuda önemli bir teknolojik kolaylaştırıcı. İnsanlığını gelişimi ve toplumsal kapsayıcılığın önündeki mevcut engelleri yıkmak için bir dizi yeni fırsat sunuyor. Yapay Zeka tarafından beslenen alanlardan biri de dil teknolojileri. Dijital asistanlar aracılığıyla telefonlarımızla iletişim kurmayı, web sitelerini ve belgeleri birkaç tıkla tercüme etmeyi, otomatik altyazı yerleştirme ile videoların erişilebilirliğini artırmayı mümkün kılan da yine dil teknolojileri.

Tüm bunların arkasındaki ana motor, **Doğal Dil İşleme (NLP)** alanındaki ilerlemeler. Peki NLP ne içeriyor? İşte bu alanın kapsamına giren temel teknolojilerin bir listesi:

Metin tabanlı:

- Makine çevirisi (*machine translation*)
- Bilgi çekme (*information retrieval*)
- Bilgi çıkarma (*information extraction*)
- Duygu analizi (*sentiment analysis*)
- Soru cevaplama (*question answering*)
- Otomatik Metin özetleme (*automatic summarization*)
- İsim verilmiş varlık tanıma (*named-entity recognition*)

Konuşma tabanlı:

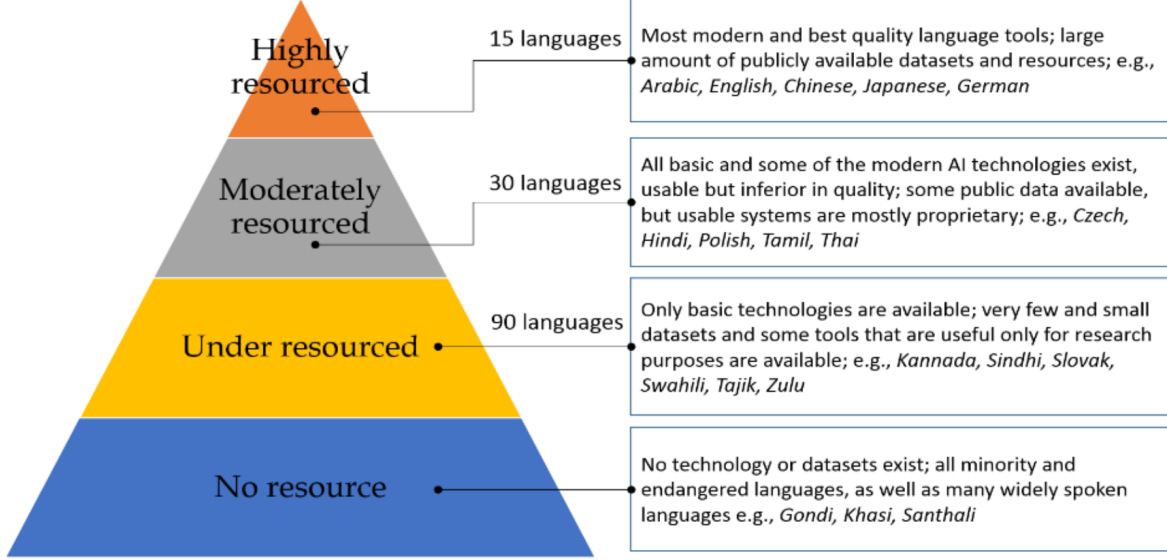
- Otomatik konuşma tanıma (*automatic speech recognition*)
- Metin-konuşma sentezi (*text-to-speech synthesis*)

Bu teknolojilerin devrim niteliğindeki yönü, *veriye dayalı* olmalarıdır. Veriye dayalı olma, bu araçlarla oluşturulan zekanın büyük hacimli bilgilerden veya daha basit ifadeyle *verilerden* toplandığı anlamına gelir. Örneğin, makine çevirisinde, motor, insan tarafından çevrilmiş belgeler ve cümleler toplamına bakarak bir dilden diğerine çeviriyi “modeller”. Benzer şekilde bir *duygu analizcisi*, insanlar tarafından iyi veya kötü bir duygu olarak etiketlenen binlerce tweet’ten yola çıkarak, bir tweet’in bir şirket hakkında iyi veya kötü bir şey söyleyip söylemediğini nasıl etiketleyeceğini öğrenir.

Bu teknolojileri bazı diller için erişilebilir kılarken diğerleri için erişilmez kılan, işte verilere olan bu bağımlılıktır. Bir dil için mevcut kaynaklar, o dil için uygulama geliştirme olasılığını doğrudan etkiler. Metinsel verilerin en büyük

kaynağı internet olduğundan ve internet birkaç dilin hakimiyetinde olduğundan, bu teknolojiler İngilizce, İspanyolca, Çince, Arapça gibi bir avuç *baskın dile* odaklanma eğilimindedir.

Microsoft Research Labs India tarafından hazırlanan aşağıdaki diyagram, diller arasındaki bu “güç yasası” tarafından yaratılan hiyerarşiyi göstermektedir.



Şekil 1: Dil teknolojilerinin, araçlarının ve kaynaklarının mevcudiyetine göre dillerin sınıflandırılması

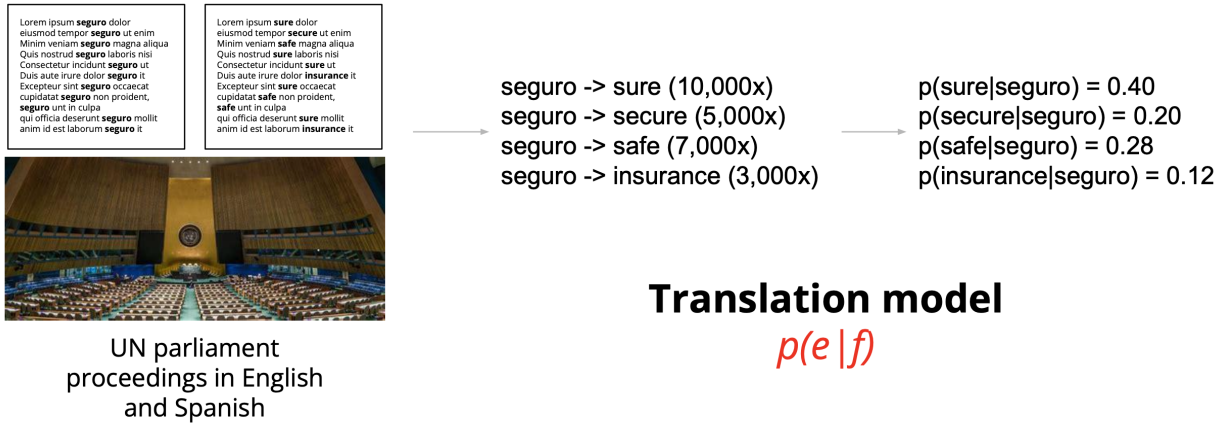
3.1 Makine çevirisi (*machine translation*)

Makine Çevirisi (MT), bir dildeki bir sembol dizisinin başka bir dilde bir sembol dizisine otomatik olarak dönüştürülmesi olarak tanımlanır. Yıllar içinde, kural bazlı yaklaşımlardan istatistiksel yaklaşımlara doğru evrilmiştir. Bu yaklaşım çeviriler arasındaki alt ifadeler arasındaki eşleme olasılıklarını modeller. Bu olasılıklar, ilgili dillerde (kaynak ve hedef diller olarak adlandırılır) cümle hizalı çevirilerin bulunduğu paralel metinlerden istatistiksel bir şekilde öğrenilir. Aşağıdaki şema, BM Parlamentosu’nda yapılan çevirileri kullanarak İngilizce’deki “*sure*” kelimesinin İspanyolca’ya çevrilmesinin modellenmesini göstermektedir.

Google Translate ve DeepL gibi makine çevirisi hizmetleri, son yıllarda çevirmenler ve çevirmen olmayan insanlar için başvuru araçları haline geldi. Bu büyük ölçüde yapay zeka alanında devrim yaratan *derin öğrenme* tekniklerinin gelişmesinden kaynaklanmaktadır. 2014 yılında tanıtılan bu yeni modelleme yöntemi, önceki modellere göre %50 daha az kelime sırası hatası, %17 daha az sözcük hatası, %19 daha az dil bilgisi hatası yaptı.

Makine çevirisinin kullanım alanları şunlardır:

1. **Benzeşme**, belirli bir belgeyi başka bir dilde taklit etme. Bu kullanım, örneğin anlamadığımız bir dilde bir haber sitesini veya teknik makaleyi okumayı sağlar. Bunun %100 doğru bir çeviri olmadığını biliriz, ancak daha fazlası için anafikri verir.
2. **İletişim**, örneğin sohbette, turizmde, e-ticarette geçer dil *lingua franca* ihtiyacını azaltarak bireyler ve kuruluşlar arasındaki iletişimi sağlar.
3. **İzleme**, büyük ölçekli çok dilli belgelerdeki bilgilerin izlenmesini sağlar, örn. Twitter’daki uluslararası trendleri keşfetmek.
4. **Yardım**, çeviri iş akışlarını iyileştirme, örn. bilgisayar destekli çeviri, post-edit.



Şekil 2: Paralel verilerden çeviri istatistiklerinin çıkarılması

MT ayrıca dil öğreniminde önemli bir araç haline geldi. Duke Üniversitesi'nin yakın tarihli bir çalışması üniversite düzeyinde dil öğrenenler arasında, sözlükler ve eş anlamlılar sözlüğü gibi diğer klasik araçların yanı sıra MT'nin kullanımı araştırıyor. İspanyolca kursunda kaydolmuş öğrencilerin %76'sının, çalışırken Google Translate gibi web tabanlı MT araçlarını kullandığını bildiriyorlar.

Son olarak, MT ayrıca Bird ve Chiang tarafından *Machine translation for language preservation* (Dil koruma için makine çevirisi) adlı makalelerinde, yok olma tehlikesindeki diller için bir belgeleme ve koruma aracı olarak önerilmiştir. Makalelerinden direkt alıntı yaparsak: "... kaynak metinler büyük bir dünya diline çevrildiğinde, dil kullanım dışı kaldıktan sonra bile dil belgelerinin yorumlanabilir olacağını garanti altına almış oluruz. İkincisi, bir dilin hala hayatta olan bir konuşanı makine çevirisi çıktılarındaki hataları tespit ettiğinde, daha fazla kapsama ihtiyaç duyan dilbilgisi ve sözlük alanlarına dair zamanında, daha fazla veri toplamak için hala fırsat varken, bilgi sahibi olmuş oluruz. Çeviri yapma veya düzeltme gibi bu işler, dışarıdan bir dilbilimcinin müdahalesine ihtiyaç duymadan o dili konuşanlar tarafından gerçekleştirilebilir. Ayrıca, oluşturulması pahalı olan ve dilin morfolojik, sözdizimsel ve semantik analizlerinin varlığına bağlı olan ağaç kümeleri (*treebank*) ve kelime ağları (*wordnet*) gibi dilsel kaynaklara duyulan ihtiyacın da önüne geçeriz."

Makine çevirisi geliştirme bu tür verilere dayandığından, bu yenilikçi dil belgeleme yöntemi, çevrilmiş cümle toplama çabasını azaltır. (Bir sonraki bölümde paralel veri hakkında daha fazla bilgi edineceğiz)

3.2 Otomatik konuşma tanıma (*automatic speech recognition*)

Otomatik konuşma tanıma (ASR), konuşmanın, akustik biçimden sözcükler veya harfler gibi bir sembolik biçime dönüştürülmesidir. "Verilen akustik girdi için, tüm olası kelime dizileri arasında en olası kelime dizisi nedir?" sorusunun olasılıksal modellemesidir. Aşağıdaki diyagram bu işlemi göstermektedir. Bir mikrofon tarafından yakalanan konuşma sinyali önce bir dizi akustik özellik vektörüne kodlanır. Ardından, bu vektörlerin kodu, konuşma sinyalinde bulunan dilsel bilgiyi temsil eden kelimelere çözülür.

Bir dil için ASR sistemi geliştirmek, aşağıdaki veri türlerine bağlıdır:

1. Birçok konuşmacıdan kısa konuşma ses kayıtlarının toplanması ve bunların yazılı transkripsiyonları
2. Büyük bir metin derlemi
3. Fonetik telaffuz sözlüğü (Bu, daha modern teknolojilerde isteğe bağlıdır)

ASR, yine derin öğrenmenin ortaya çıkması sayesinde, son on yılda önemli ölçüde ilerlemiştir. Eylül 2017'de Microsoft, bir konuşmanın deşifresinde insandan daha iyi performans elde edebilecek bir İngilizce konuşma tanıma sistemine dair



Şekil 3: Otomatik konuşma tanımının basit bir diyagramı

sonuçlarını açıkladı. Microsoft'un sistemi, günlük konuşmalardan deşifre edilmiş 200 milyon kelimeden oluşan bir veri kümesine dayanıyordu. Bu gelişmeler, sanal asistanların her gün kullanılan bir uygulama, sesli arama ve sesin otomatik transkripsiyonu haline gelmesiyle şimdiden büyük etki yarattı.

3.3 Metin-konuşma sentezi

Konuşma sentezi (TTS), bir bilgisayarla, verilen metin girişi için insan benzeri bir konuşmanın üretilmesini içerir. Derin öğrenmenin ortaya çıkmasından önce, TTS'e yönelik iki ana yaklaşım vardı: birleştirici TTS ve parametrik TTS. Birim seçimi olarak da adlandırılan birleştirici TTS, istenen metni sentezlemek için birimler adı verilen önceden kaydedilmiş kısa ses kliplerini birleştirir. Birleştirici TTS, konuşma kalitesi açısından iyi bir performans sağlayabilir ancak kes-birleştir işlemi, doğallıktan yoksun olduğu anlamına gelir. Parametrik TTS, insan konuşma oluşumunu modelleyerek F0 (temel frekans) ve enerji gibi parametrelerin bir kombinasyonu ile konuşma üreten istatistiksel yöntemlere dayanır.

Şu anda, çoğu modern TTS sistemi, derin öğrenme yöntemlerine dayanır. Derin sinir ağları, büyük miktarlarda kaydedilmiş konuşma ve karşılık gelen metin deşifreleri kullanılarak eğitilir. Bu konuşmalar, ASR eğitim verilerinin aksine, genellikle tek bir konuşmacıdan toplanır. Ortaya çıkan TTS sistemi, bu belirli konuşmacının sesini kopyalama becerisine sahiptir.

TTS, ekrandan "okuma"yı mümkün kıldığı için görme engelli veya kısmen gören kişiler için bilgisayarları erişilebilir hale getirmede önemlidir. TTS teknolojisi, çeşitli dillerde herhangi bir yazılı girdiye bağlanabilir, örn. bir çevrimiçi sözlükte kelimelerin otomatik telaffuzu, bir metnin sesli okunması, sesli asistan için arayüz vb.

Tehlike altındaki diller ve azınlık dilleri söz konusu olduğunda, TTS dil öğrenimine ve dil dokümantasyonuna yardımcı olabilir. Söz konusu dili konuşan kişilere erişimi olmayan öğrenciler, bir öğretmenin yardımı olmadan bir cümlenin nasıl telaffuz edildiğini öğrenebilirler. Dilin kalıcı bir kayıttır, zira o dil için hiç konuşmacı kalmadığı andan sonra bile varlığını sürdürecektir.



Şekil 4: Bu yayın Avrupa Birliği'nin maddi desteği ile hazırlanmıştır. İçerik tamamıyla Col-lectivaT ve SKAD'ın sorumluluğundadır ve Avrupa Birliği'nin görüşlerini yansıtmak zorunda değildir.



Şekil 5: Bu proje Avrupa Birliği tarafından finanse edilmiştir.

Yapay zeka (AI) araçları, tehlike altındaki diller ve azınlık dilleri için dil kaynağı oluşturma adına yeni bir alan açar. Sözlük, dil bilgisi belgeleri, dil haritaları vb. gibi dilleri korumak için oluşturulan “klasik” dil kaynaklarıyla karşılaştırıldığında, AI araçları daha az dilbilimsel uzmanlık gerektirir, ancak genellikle yalnızca büyük hacimlerde olduğunda yararlıdır.

Bu bölümde, bir önceki bölümde anlattığımız yapay zeka tabanlı dil teknolojilerinin oluşturulmasını besleyen veri türlerini açıklayacağız. Ayrıca, genellikle bu uygulamaların gerektirdiği gibi büyük hacimli olmasalar bile, bu verileri toplamının ve bunlardan en iyi şekilde yararlanmanın bazı yollarını ele alacağız.

4.1 Metin derlemi

Dilbilimde, bir metin derlemi (İng. *text corpus/corpora*), bir dilde dijital formatta büyük ve yapılandırılmış bir metin kümesinden oluşan bir dil kaynağıdır. Derlem dilbiliminde (*corpus linguistics*) metin derlemlerinden, istatistiksel analiz yapmak ve hipotezi test etmek, oluşumları kontrol etmek veya belirli bir dil bölgesi içindeki dil kurallarını doğrulamak için faydalanılır. Dil teknolojileri açısından, optik karakter tanıma, el yazısı tanıma, makine çevirisi, yazım hataları düzeltme, bilgisayar destekli yazım gibi uygulamalarda kullanılan istatistiksel dil modellerinin oluşturulmasında önemli bir rol oynarlar.

Metin derlemleri kendi başlarına *etiketlenmemiş veri* türündedir. Yani, sadece bir veri yığındır (bu durumda metin), herhangi bir açıklama veya etiket içermez. Dil modelleri, dille ilgili bir “fikir edinmek” için kelime dizilerinin olasılıklarını depolar. Ayrıca, farklı NLP görevleri için *etiketli veriler* oluşturmak amacıyla metin derlemlerine aşağıdaki açıklamalar eklenebilir:

- **Sözcük türü** (İsim, fiil, Sıfat vb.)
- **Adlandırılmış varlıklar** (Kişi, konum, kişisel olarak tanımlanabilir bilgiler, kurum, zaman vb.)
- **Lemma** (kelime kökleri, ör. kırdı için kırmak gibi)
- **Bağımlılık ve öbek yapısı** (sözdizimi ağacı)

4.1.1 Metin derlemelerine kaynak sağlama

Metin derlemesine kaynak sağlamanın en yaygın yolu, web'i *taramaktır*. Bu teknik, aynı anda belirli bir dilde veya birçok dilde metin toplamak için tüm web'i ayrıştırır (*parse*). Örneğin Wikipedia, içeriğini farklı dillerde [yayınlar](#), bu da bir metin derlemi oluşturmak için kullanılabilir. [Common Crawl](#) inisiyatifi, web sitesi verilerini toplar ve halka petabaytlar (1 PB=1024 terabayt) veriyi ücretsiz olarak sağlar. OSCAR, bu verileri 166 dilde sınıflandırarak dağıtır.

Kaynağı olan diller için kullanılan bir diğer yaygın kaynak ise kitaplardır. [BookCorpus](#) web'den 74 Milyon cümle ve 984 Milyon kelime içeren 11.038 kitaptan oluşuyor ve büyük teknoloji şirketlerinin oluşturduğu birçok etki yaratan dil modelini desteklediği biliniyor.

Not: Web'de ve kitaplardaki dil/üslup önyargı ya da toksik dil içerir. Buralarda açıkta bulunan verilerden oluşturulan dil modelleri, gördüklerini temsil etmekten başka bir şey yapmaz. Dolayısıyla bu üsluplar, modellerde yeniden üretilir. Büyük dil derlemlerinden dil modelleri oluşturmanın olası risklerinin analizi için, [Bender et al. tarafından yazılan bu makaleye](#) bakabilirsiniz.

4.2 Paralel veri (bibtex)

Paralel veri, bir makine çevirisi sistemi oluşturmak için ihtiyaç duyulan veri türüdür ve bir dildeki cümlelerin çevirileriyle birlikte bir araya gelmesinden oluşur. Tarihsel olarak paralel verilerin kaynağı Birleşmiş Milletler ya da Avrupa Parlamentosu gibi çok dilli kamusal alanlarda yapılan çeviriler olmuştur. Şu anda ise paralel metnin en büyük kaynağı çok dilli web'dir.

Makine çevirisi modellerini eğitmek için sadece tercüme edilmiş belgelere sahip olmak yeterli değildir. Metinlerin cümlelere bölünmesi ve hizalanması gerekir. Paralel metin hizalama, paralel metnin her iki tarafında birbirine karşılık gelen cümlelerin tanımlanmasıdır. Ortaya çıkan belgeler ya satır satır karşılık gelmeli ya da orijinal cümleleri ve çevirilerini aynı satırda içermelidir. [Hunalign](#) çevrilmiş belgelerden cümle hizalamaları oluşturmaya yardımcı olur. TMX uzantılı çeviri hafızası dosyaları (*Translation Memories*) da hali hazırda cümle cümle segmentlere ayrılmış olduğu için iyi paralel veriler oluşturur.

4.2.1 Paralel veriye kaynak sağlama

[OPUS](#) hemen hemen tüm kamuya açık paralel verilerin bir derlemesidir. Birçok araştırmacının, MT modelleri için kaynak verilerini toplamak veya paralel verilerini yayınlamak için başvurduğu ilk yerdir.

Paralel veriler için bazı yaygın kaynaklar şunlardır: Çok dilli web siteleri (ör. uluslararası haber kuruluşları), film alt-yazıları (bkz. [OpenSubtitles](#)), kutsal metinler, parlamento oturumları, yazılım yerelleştirme verileri.

4.2.2 Tatoeba.org ile kitle kaynaklı paralel veriler sağlama

Tatoeba, yabancı dil öğrenenlere yönelik hazırlanmış, çevirileri ile birlikte sunulan ücretsiz bir örnek cümleler veri tabanıdır. Açık ve ortak çalışma modeli aracılığıyla bir gönüllüler topluluğu tarafından yazılır ve sürdürülür. Bağışlarla finanse edilen kar amacı gütmeyen Fransız kuruluş Tatoeba Association tarafından sunulur. Şu anda 412 desteklenen dilde 10.397.308 cümle barındırır.

Kullanıcılar, herhangi bir dilde kelimeleri arayabilir, bu kelimelerin geçtiği cümleleri bulabilirler. Tatoeba veri tabanındaki her cümle, yanında diğer dillerdeki olası çevirileriyle birlikte gösterilir; doğrudan ve dolaylı çeviriler birbirinden ayrılır. Cümle içerikleri konu, diyalekt veya müstecenlik gibi etkileşimlerle belirtilir; ayrıca kültürel notlar ve diğer kullanıcılardan gelen geri bildirimler ve düzeltmeleri kolaylaştırmak için her birinin ayrı yorum zincirleri bulunur. Cümleler dil, etiket ve başka kriterlere göre aranabilir.

Random sentence

Sentence #1533839 — belongs to GrizaLeono



Êu vi scias, kiu verkis tiun romanon?





Translations

>



Weißt du, wer diesen Roman verfasst hat?





>



Μήπως γνωρίζετε ποιος έγραψε αυτό το μυθιστόρημα;





>



Sais-tu qui a écrit ce roman ?





Translations of translations

>



Tezriđ anwa i yuran ungal-a ?





>



Weißt du, wer diesen Roman geschrieben hat?





SHOW 7 MORE TRANSLATIONS

Şekil 1: Tatoeba'dan bir cümle ve çevirileri

Kayıtlı kullanıcılar, hedef dil kendi ana dilleri olmasa bile yeni cümleler ekleyebilir veya mevcut cümleleri çevirebilir veya düzeltebilir. Ancak, kullanıcıların kendi ana dillerinde veya en iyi bildikleri dillerde orijinal cümleler veya çeviriler eklemeleri önerilir.

Tatoeba veritabanının tamamı Creative Commons Atıf 2.0 lisansı altında yayınlanmaktadır. Ayrıca, [downloads sayfasından](#) parça parça derlemleri tek dilli veya paralel biçimde indirmek çok kolaydır.

4.3 Konuşma derlemi

Bir konuşma derlemi (*speech corpus/corpora*), konuşma ses dosyalarının ve genellikle bunların metin transkripsiyonlarının bir toplamıdır. Konuşma teknolojisinde konuşma derlemleri otomatik konuşma tanıma, metin-konuşma sentezi veya konuşmacı tanımlama gibi görevler için akustik modeller oluşturmak için kullanılır.

Konuşma derlemi okunmuş (ör. sesli kitaplar, haberler, okunmuş sayı veya kelimeler) veya doğal konuşmalar (diyaloglar) içerebilir. ASR (otomatik konuşma tanıma) modelleri eğitmeye uygun derlemler, mümkün olduğunca çok konuşmacıdan çeşitli akustik ortamlarda (ör. gürültülü, uzaktan) toplanan örnek konuşmaları içerir. Bunun tersine, TTS eğitim verileri çoğu zaman ideal akustik ortamda tek bir konuşmacıdan alınan kayıtlardan oluşur.

[OpenSLR](#) halka açık birçok konuşma derlemini listeler.

4.3.1 Common Voice

[Common Voice](#), konuşma tanımayı herkes için erişilebilir kılmak için ücretsiz bir veritabanı oluşturmak amacıyla Mozilla tarafından başlatılan bir kitle kaynaklı projedir. Proje, mikrofonla örnek cümleler kaydeden ve diğer kullanıcıların kayıtlarını kontrol eden gönüllüler tarafından desteklenir. Seslendirilmiş örnekler, [CC0 kamu malı \(public domain\)](#) lisansı altında düzenli aralıklarla yayınlanır. Bu lisans, geliştiricilerin veritabanını sestene-metne-çevirme uygulamaları için kısıtlama veya maliyet olmaksızın kullanabilmelerini sağlar.

Not: Mayıs 2022 itibarıyla Common Voice 63 dili desteklemektedir, 68 yeni dil de desteklenme aşamasındadır. Mevcut dil listesine [buradan](#) bakabilirsiniz.

4.3.2 Common Voice’a dil ekleme

Common Voice, her dilin kendi topluluğuna sahip olduğu bir topluluk platformu olarak çalışır. Common Voice’a yeni bir dilin eklenmesi prosedürü aşağıdaki gibidir:

1. **Dil için bir topluluk yöneticisi bulma** ([Roller hakkında bilgi](#))
2. **Mozilla’ya yerleştirme isteği gönderme** Bu, github sayfalarında [bu şablon](#) kullanılarak yapılır. Bu işlem, istenen dili Pontoon’a yerleştirmek suretiyle Common Voice’un bu dile yerleştirme sürecini başlatacaktır.
3. **Pontoon’da yerleştirme süreci** ([kullanım kılavuzu](#)). Common Voice platformundaki her dize, stil kılavuzuna uygun olarak yerleştirilmek istenen dile çevrilmelidir. Toplam 663 dize bulunur. Çeviriler, platforma kaydolun ve o dili konuşan herhangi bir kullanıcı tarafından yapılabilir, ancak bu çevirilerin topluluk yöneticisi tarafından gözden geçirilmesi gerekir.
4. **Cümle toplama** En az 5000 kamu malı olan cümlelerin toplanması ve [Common Voice cümle toplayıcıya](#) girilmesi gerekir.
5. **Cümlelerin gözden geçirilmesi** Toplanan her cümle, cümle toplayıcıda en az iki kullanıcı tarafından manuel olarak kontrol edilmelidir.



Şekil 2: Common Voice'ta Kurmanci Kürtçe bir cümle kaydetme

6. **Bir sonraki Common Voice sürümünü bekleme** Yerelleştirme tamamlandığında ve 5000 cümlelerin kontrolü bittiğinde, bir sonraki Common Voice sürümü dilinizi içerecektir. Sürümler ayda iki kez yapılır ve takvimleri [github verihavuzunda](#) listelenir.

4.3.3 Bulunmuş veriler

Ayrıca, radyo programlarından, filmlerden ve röportajlar gibi kayıtlı materyallerden ses verileri elde etmek de mümkündür. Bu tür veriler, orijinal olarak ses teknolojisi oluşturmaya hizmet etmesi amaçlanmadığı için “bulunmuş veriler” olarak adlandırılır, ancak *başka bir amaca hizmet* etmek üzere uyarlanır. Bulunmuş verilerin kısa ses segmentleri ve bunların transkripsiyonlarını elde etmek için işlenmesi gerekir.

4.3.4 Kaynaklar

- Wikipedia’da metin derlemi
- Wikipedia’da Tatoeba
- Wikipedia’da konuşma derlemi
- Wikipedia’da Common Voice



Şekil 3: Bu yayın Avrupa Birliğinin maddi desteği ile hazırlanmıştır. İçerik tamamıyla Col-lectivaT ve SKAD’ın sorumluluğundadır ve Avrupa Birliğinin görüşlerini yansıtmak zorunda değildir.



Şekil 4: Bu proje Avrupa Birliği tarafından finanse edilmiştir.

Örnek çalışmalar

Bu bölümde, yok olma tehlikesi altındaki diller ve azınlık dilleri için örnek veya ilham verici olduğunu düşündüğümüz dil teknolojisi ile ilgili bazı girişimleri ve çalışmaları listeliyoruz. Bu belgenin de bünyesinde hazırlandığı “Judeoespanyol: Akdeniz’in iki yakasını birleştiren dil” projesini anlatarak başlayıp Anadolu, İber, Afrika dillerini içeren projelerle devam edeceğiz.

5.1 Judeoespanyol: Akdeniz’in iki yakasını birleştiren dil

Col-lectivaT ve Sefarad Kültür Araştırma Merkezi (SKAD), sosyal medya için dil eğitimi içeriklerinin oluşturulmasından, Ladino’ya dijital çağa yardımcı olacak ileri dil teknolojilerinin geliştirilmesine kadar çok çeşitli faaliyetler gerçekleştirmek üzere bu proje için bir araya geldi. Ayrıca bu dilin Türkiye ile İspanya arasında ortak bir kültür mirası olduğu konusunda farkındalık oluşturmayı amaçlıyor.

5.1.1 Sosyal medya için görsel-işitsel içerik oluşturulması

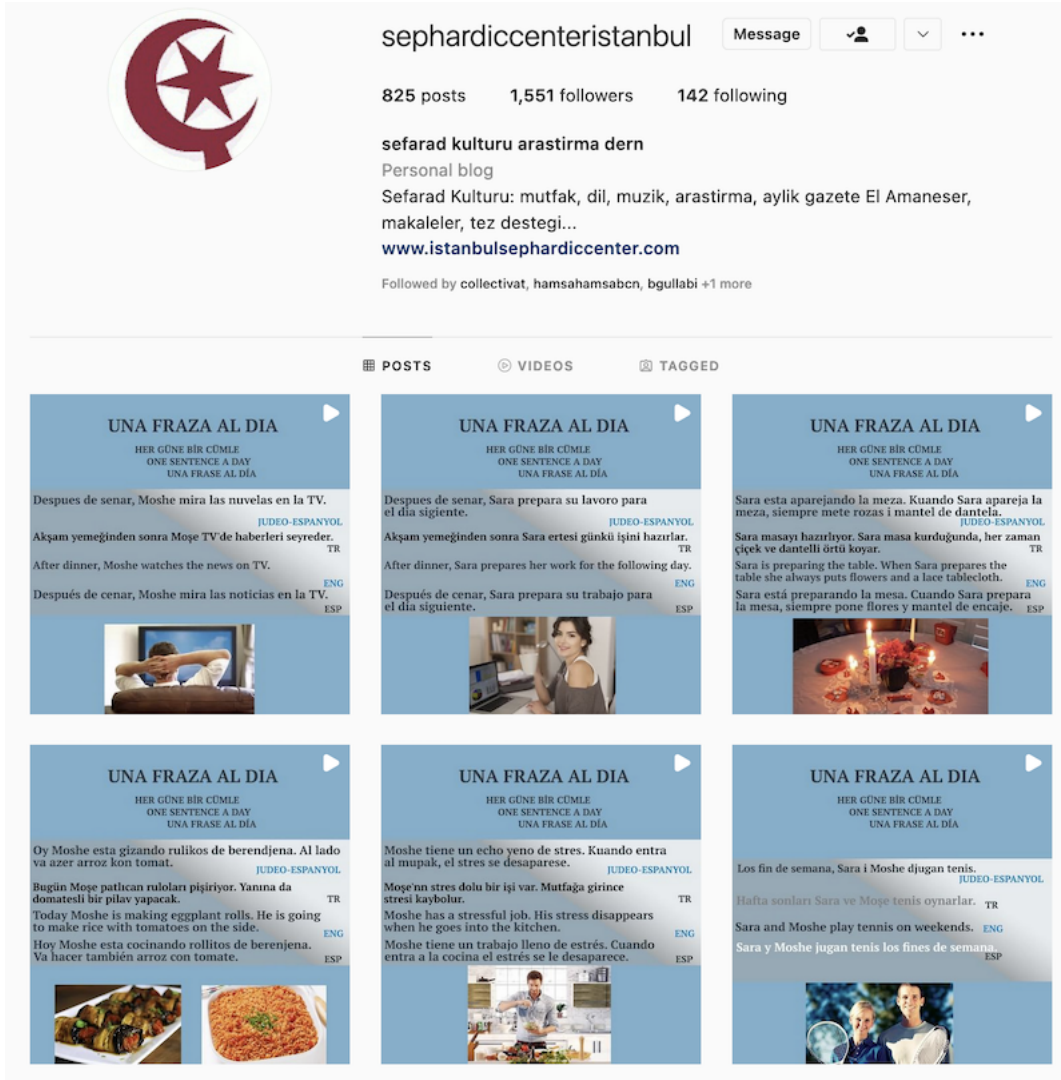
Proje, sosyal medya platformlarında görünürlük kazanmaya yardımcı olacak ve genç nesilleri Ladino öğrenmeye çekecek kısa dil öğrenme videoları oluşturdu. Bu kısa videolarda bir Ladino cümleyi Türkçe, İngilizce ve İspanyolca çevirisi ile telaffuzunu öğrenmeye yardımcı olacak sesli bir şekilde sunuluyor.

5.1.2 Ladino Data Hub ve açık dil veri kümeleri

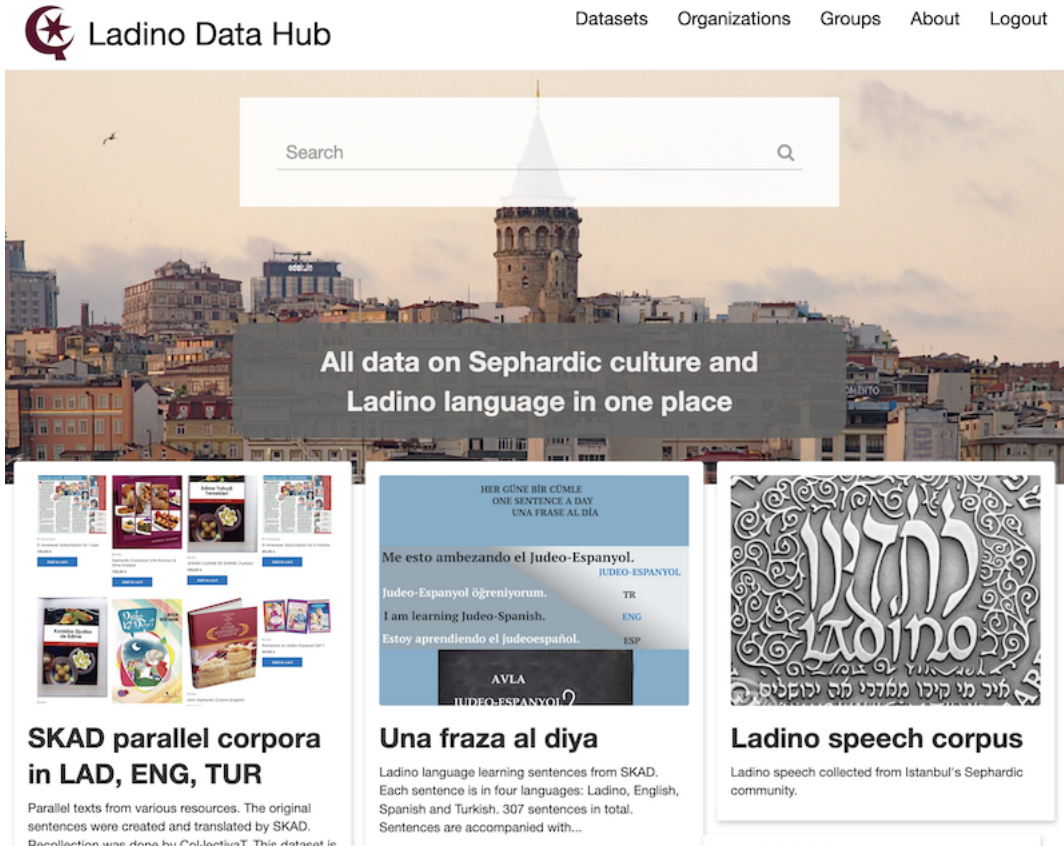
Proje, Ladino dil verilerini ve Sefarad kültürünün belgelenmesine yardımcı olan diğer kaynakları barındırmaya ayrılmış merkezi bir web arşivi görevi görecektir olan [Ladino Data Hub](#)’ı oluşturdu. Dünyanın her yerinden araştırmacıların, gazetecilerin Ladino için araştırma ve geliştirmeyi arttırmaya yardımcı olacak veri kümelerine erişmesini ve bunları paylaşmasını sağlamayı amaçlıyor.

Proje, halihazırda var olan veri kümelerini yeniden paketledi, yeni veri kümeleri oluşturdu ve bu portalda paylaştı. Bunlar:

- [Şalom gazetesinden](#) taranan cümlelerden oluşan bir [metin derlemi](#)



Şekil 1: Una Frazza al diya (Her gün bir cümle) videoları ile SKAD'ın Instagram sayfasında Ladino tanıtımı

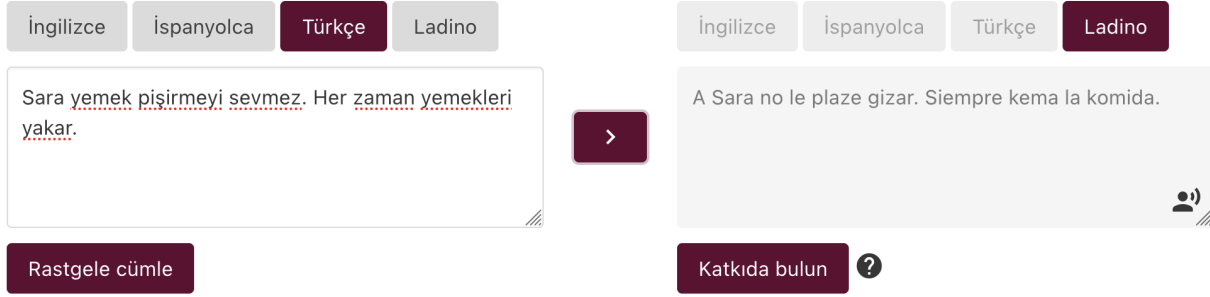


Şekil 2: Ladino Data Hub, Ladino dili ve Sefarad kültürüyle ilgili verileri barındırıyor

- Una Frazza al diya cümlelerini ve ses kayıtlarını içeren sesli paralel metin derlemi
- SKAD tarafından oluşturulan Ladino konuşma derlemi
- Karen Şarhon tarafından okunan konuşma materyallerinden oluşan temiz konuşma sentezi eğitimi veri seti
- Güler'in sözlüğüne dayanan İspanyolca Ladino sözlüğü (lexicon)
- SKAD bünyesinde yapılan çevirilerden `Ladino, İngilizce ve Türkçe paralel metinler
- Temel makine çevirisi modellerinin eğitimi için sentetik olarak üretilmiş paralel derlem

5.1.3 Makine çevirisi ve konuşma sentezi için web uygulaması

Projenin son ve en önemli çıktısı, Ladino ile Türkçe, İspanyolca ve İngilizce arasında çeviri yapabilen bir web uygulamasıdır. Amaç, Ladino öğrenmek isteyen kişilere, araştırmacılara ve dilbilimcilere yardımcı olmaktır. Makine çevirisi back-endi, [kural tabanlı bir makine çeviri sistemi](#) yardımıyla oluşturuldu. Bu, İspanyolca ve Ladino arasındaki benzer sözdiziminden faydalanıp, ancak sözlüklerden ve dilbilgisi kitaplarından türetilen bir dizi kuralla yazım ve kelime dağarcığını değiştirerek İspanyolca'dan LAdino'ya dönüştürebilen bir arkayüz. Web uygulaması ayrıca [TTS eğitimi veri kümesi](#) ile oluşturulmuş bir metin konuşma sentezi uygulaması ile Ladino cümleleri sentezleyebiliyor.



Şekil 3: Konuşma sentezi ile Ladino çeviri web uygulaması

Web uygulaması ayrıca paralel verilere katkıda bulunmaya da izin verir. Kullanıcılar, Ladino için paralel verileri genişletmek için rastgele bir cümle yükleyebilir ve düzeltilmiş çevirilerini gönderebilir.

Not: Bu projenin ayrıntılı bir teknik raporu için Avrasya'daki Yerli, Tehlikedeki ve Az Kaynaklı Diller için Kaynaklar ve Teknolojiler Çalıştayında ([EURALI](#)) sunulan “Tehlike Altındaki Bir Dili Dijital Çağa Hazırlamak: Judeoespanyol Örneği” başlıklı makaleye bakabilirsiniz: [Bağlantı yakında eklenecek](#)

5.2 Tabandan örgütlenen NLP toplulukları

Dikkatlerini sadece birkaç dile veren NLP araştırmalarına karşı, tabandan gelen araştırma toplulukları, dünyadaki dilleri teknolojinin ön saflarına getirmek için bir araya geliyor. Bu girişimlere iki örnek [Masakhane](#) ve [Turkic Interlingua](#).

Web sayfalarında tanımlandığı gibi, “Masakhane, misyonu Afrikalılar tarafından Afrikalılar için Afrika dillerinde NLP araştırmalarını güçlendirmek ve teşvik etmek olan bir taban örgütüdür.” Bu, adlarından da anlaşılacağı gibi “birlikte inşa etmek” için herkese açık bir girişimdir. Dil teknolojisi araştırmalarında Afrika'nın 2000 dilini temsil etmek amacıyla tüm dünyada Afrikalı ve Afrikalı olmayan araştırmacılar tarafından birçok eş zamanlı etkinlik gerçekleştirilmektedir. Masakhane'nin öne çıkan bazı çalışmaları:

- **Masakhane çeviri** 6 Afrika dilini destekliyor: Yoruba, Shona, Lingala, Swahili, Tshiluba
- **Lanfrica** Afrika dil eserlerini bulmada karşılaşılan zorluklara karşı koymak için Afrika dil kaynaklarını katalogluyor
- **BibleTTS** Sahra Altı Afrika’da konuşulan on dil için yüksek kaliteli Metin Okuma modellerinin geliştirilmesini sağlıyor: Ewe, Hausa, Kikuyu, Lingala, Luganda, Luo, Chichewa, Akuapem Twi, Asante Twi, Yoruba.
- **Oshiwambo dilini ve kültürünü korumak için makine çevirisi**
- **MasakhaNER Know our names** (MasakhaNER İsimlerimizi bilin), çeşitli Afrika dilleri için elle, adlandırılmış varlık tanıma (NER) veri kümeleri oluşturma

Masakhane ayrıca Afrika NLP ile ilgili araştırmaları yayınlamak için yıllık çalıştaylar düzenliyor ve **Lacuna** gibi veri toplama fonlarına katılıyor.

Türkic Interlingua (TIL), Altayca, Azerice, Başkurtça, Şorca, Kırım Tatarcası, Çuvaşça, Gagauzca, Karakalpakça, Hakasça, Kazakça, Karay-Balkarca, Kumukça, Kırgızca, Sahaca (Yakutça), Salarca, Türkmençe, Türkçe, Tatarca, Tuvaca, Uygurca, Urumca ve Özbekçe gibi “Türkî diller için, misyonu yazım denetleyicilerinden çeviri modellerine kadar dil teknolojileri geliştirmek, çeşitli veri kümeleri toplamak ve dilbilimsel olguları akademik araştırma merceğinden incelemek olan bir araştırmacılar, mühendisler, dil meraklıları ve topluluk liderlerinden oluşan bir topluluktur.”

5.3 Common Voice kampanyaları

Çeşitli dil toplulukları, Common Voice’a katılımı harekete geçirmeye başladı. Bu girişimler, tek tek bireylerden yerel yönetimlere kadar uzanan gruplar tarafından oluşturuluyor. Bazı örnekler:

- **Common Voice Türkçe**
- **Kurdish crowdsorce**
- **Igbo dili için Igwebuike topluluğu**
- **Katalanca için AINA projesi**

Common Voice’a yaptıkları katkılardan dolayı Katalan halkından özel olarak bahsetmek gerek. İspanya’da vatansız bir azınlıklaştırılmış dil olan Katalanca, aktivistlerin inanılmaz katkıları ve ayrıca **Katalan yerel hükümetinin AI girişimi** sayesinde Mayıs 2022 itibarıyla Common Voice’taki 4.büyük dil.

5.4 Diğer girişimler

Bahsetmeye değer diğer bazı girişimler şunlar:

- **Col·lectivaT**, Katalan kamu televizyon yayınlarını ve meclis oturumları kayıtlarını kullanarak **Katalanca açık kaynak konuşma verileri ve metin derlemleri** oluşturdu.
- **Catotron**, Katalunya Kültür Bakanlığı’nın desteğiyle oluşturulmuş Katalanca ilk açık kaynak, sinir ağı (*neural network*) tabanlı konuşma sentezi motorudur.
- **Filipinler’in** dilleri Cebuano ve Waray-Waray, makine çevirisinin rekabetçi kullanımı sayesinde en büyük Wikipedia sayfalarından birine sahiptir (*kaynak*)
- **Maori halkı**, “kendi kaderini tayin hakkını korumak” için kendi dillerinde ses verileri toplamaya yönelik özel veya açık kaynaklı girişimleri reddetti (*kaynak*)
- **Açık Dil Teknolojisi Manifestosu**
- **Hindistan Microsoft Research Tarafından ELLORA Düşük Kaynaklı Dilleri Etkinleştiriyor**



Şekil 4: New York Times Meydanı'nda "İnternetin Katalanca konuşmasının zamanı geldi" yazan bir reklam panosu (fotoğraf: Aina Martí)



Şekil 5: Bu yayın Avrupa Birliği'nin maddi desteği ile hazırlanmıştır. İçerik tamamıyla Col-lectivaT ve SKAD'ın sorumluluğundadır ve Avrupa Birliği'nin görüşlerini yansıtmak zorunda değildir.



Şekil 6: Bu proje Avrupa Birliği tarafından finanse edilmiştir.

İyi uygulamalar

Not: Bu bölüm, proje kapsamında düzenlenen çalıştaylarda toplanan geri bildirimlerle geliştirilmeye devam edecektir.

Dünyadaki çok sayıda dil için teknolojiyi etkinleştirmek üzere kaynak yaratmanın yoğun yatırım gerektirdiği açık olsa da, böyle bir yatırımın kısa sürede hızlıca ve kolayca gerçekleşmesi pek olası görünmüyor. Bu sınırlı kaynaklar göz önüne alındığında, dil toplulukları, dillerinin geleceğini belirleyebilmek için güçlendirilmediler. Bu belgede, dijital temsilin bu sürecin nasıl bir parçası olduğunu anlattık.

Nereden başlayacağınıza karar vermek için, Bali et al. tarafından **ELLORA** inisiyatifinde tanıtıldığı gibi, *Keşfet, Tasarla, Geliştir ve Etkinleştir* anlamına gelen 4-D (*Discover, Design, Develop and Deploy*) tasarım odaklı düşünme yöntemini benimsemenizi öneriyoruz. Bu kullanıcı merkezli yaklaşım aşağıdaki gibidir:

1. Dil topluluğu tarafından en çok neye ihtiyaç duyulduğunu keşfedin,
2. Teknolojiyi, dilin çeşitliliğine dikkat ederek ve bir çoğunluk dilinden yola çıkan bir yaklaşımdan kaçınarak, kullanıcılar ve dilleri için tasarlayın,
3. Hataları baştan itibaren sürekli tespit ederek ve iyileştirerek interaktif bir şekilde teknolojiyi sık sık geliştirin ve etkinleştirin.

Topluluk, henüz bir dil teknolojisi geliştirme perspektifine sahip değilse bile, dil koruma etkinlikleri düzenlerken verilerin değerini akılda tutmak iyi uygulamalara girer. Bazı örnekler şunlardır:

- Dil topluluğunda veri farkındalığını artırmak için etkinlikler düzenlemek ve içerik oluşturmak,
- Dilleri kitle kaynaklı çalışma (*crowdsourcing*) platformlarında tanıtmak,
- Dil verilerinin toplanması için veri toplama maratonları (*datathon*) düzenleyin,
- Kamuya açık hale gelen halk hikâyelerini ve masalları tercüme edin,
- Metin derlemi (*text corpora*) oluşturmak için yayınlanmış materyalin düz metin (*plain text*) veya belge versiyonlarını saklayın,
- Diğer çevirmenlere yardımcı olmak ve paralel veriler oluşturmak için, çeviri belleklerini (TM) kaydedin ve açık olarak paylaşın,

- Yayın materyallerinin (örneğin radyo programları) kayıtlarını saklayın ve mümkünse konuşma verilerine dönüştürülebilmeleri için metine aktarın-deşifre edin,
- Sosyal medyada yayınlanan içerikleri, zaman tünellerinde kaybolmamaları için kalıcı bir yere kaydedin.

Önerileriniz veya sorularınız var mı? info-at-collectivat.cat adresinden bize yazın



Şekil 1: Bu yayın Avrupa Birliği'nin maddi desteği ile hazırlanmıştır. İçerik tamamıyla Col-lectivaT ve SKAD'ın sorumluluğundadır ve Avrupa Birliği'nin görüşlerini yansıtmak zorunda değildir.

BÖLÜM 7

Lisans

Bu belge, Attribution 4.0 International (CC BY 4.0) lisansı ile lisanslanmıştır.



Şekil 1: Bu yayın Avrupa Birliğinin maddi desteği ile hazırlanmıştır. İçerik tamamıyla Col-lectivaT ve SKAD'ın sorumluluğundadır ve Avrupa Birliği'nin görüşlerini yansıtmak zorunda değildir.